A Unified Framework for Defining and Identifying Causal Effects

Halbert White and Karim Chalak
Department of Economics
University of California, San Diego
La Jolla, CA 92093-0508
USA

January 30, 2006

Abstract This paper unifies three complementary approaches to defining, identifying, and estimating causal effects: the classical structural equations approach of the Cowles Commision; the treatment effects framework of Rubin (1974) and Rosenbaum and Rubin (1983); and the Directed Acyclic Graph (DAG) approach of Pearl. The settable system framework nests these prior approaches, while affording significant improvements to each. For example, the settable system approach permits identification of causal effects without requiring exogenous instruments; instead, a weaker conditional exogeneity condition suffices. It removes the stable unit treatment value assumption of the treatment effect approach and provides significant insight into the selection of covariates. It generalizes the DAG approach by accommodating mutual causality and attributes. We provide a variety of results ensuring structural identification of general covariate-conditioned average causal effects, laying the foundation for parametric and nonparametric estimation of effects of interest and new tests for structural identification.

Acknowledgments: The authors are indebted to Jinyong Hahn, James Heckman, Kei Hirano, Kevin Hoover, Meng Huang, Massimiliano Marinucci, Rosa Matzkin, and Rosalind Wu for helpful discussions and suggestions. Any errors remain the authors' responsibility

JEL Classificiation Numbers: .

1 Introduction

The introduction by Tinbergen, Frisch, Koopmans, Haavelmo, and the other pioneers of the Cowles Commission in the late 1940s and early 1950s of the "simultaneous equations" or "structural equations" approach to econometric modeling revolutionized economics by providing sophisticated methods with which to attempt to understand and measure causal effects in economics. Despite this momentous advance, economists' focus on causal relationships declined in the following decades, due at least in part to the difficult conceptual and statistical issues raised in the ensuing scholarly debate. Hoover (2004) provides enlightening documentation and discussion of this decline; he further documents the strong resurgence of interest in causal issues in economics over the last decade.

This resurgence has been significantly fueled by the work of a group prominently populated by labor economists, including that of Heckman and Robb (1985), Heckman, Ichimura, and Todd (1997, 1998), Angrist (1998), Hahn (1998), Hirano and Imbens (2001), Hirano, Imbens, and Ridder (2003), Heckman, Urzua, and Vytlacil (2005) (HUV), and Heckman and Vytlacil (2005) (HV). These authors have developed powerful new methods for estimating causal effects, blending the classical structural equations approach with methods developed in the treatment effects literature by Rubin (1974) and Rosenbaum and Rubin (1983).

Another important strand of thinking about causal structures has emerged from the artificial intelligence literature. There, Pearl (1988, 1993a, 1993b, 1995, 1998, 2000) and his colleagues (Spirtes, Glymour, and Scheines, 1993 (SGS) and Dawid, 2002, among others) have developed insightful techniques for analyzing causal relations using directed acyclic graphs (DAGs).

These different approaches to understanding causality are not fully compatible, as each admits certain features ruled out by the others. So far, there does not exist a formal mathematical framework that encompasses these various approaches to defining, identifying, modeling, and estimating causal effects. Our goal here is to provide such a framework, permitting rigorous definitions of notions of cause and effect; and then, based on this framework, to provide conditions for the identification of causal effects of interest. In a companion paper (White and Chalak 2006), we analyze the statistical properties of parametric and nonparametric estimators of these effects, and provide new tests for identification of causal effects.

The "settable system" framework proposed here requires a non-trivial reconsideration of the foundations of each of the approaches it seeks to unify. This would not be worth the required effort unless it results in significant benefits relative to each of these prior approaches. In fact, settable systems offer important improvements and advantages in each case.

Specifically, relative to the classical structural equations approach, settable systems permit the identification and estimation of well-defined causal effects in the absence of traditional exogenous variables. Instead, a weaker conditional exogeneity condition suffices, yielding an extension of the concept of instrumental variables. Our framework shares the flexibility of modern treatments of structural equations, such as Matzkin (2003, 2004, 2005), Imbens and Newey (2003), HUV, and HV, by not imposing the classical restrictions of linearity or separability on the structural relations. The effects of interest identified here do not, however, require the monotonicity requirements imposed by Matzkin (2005) and Imbens and Newey (2003). Our framework also accommodates unobserved ("essential") heterogeneity emphasized by HUV and HV. It thus provides a framework complementing the work of Matzkin (2003, 2004, 2005), Imbens and Newey (2003), HUV, and HV that permits the causal notions either implicit or explicit in their approaches to be fully formalized and that permits generalizations of their methods. The identification and estimation results that emerge also provide a significant refinement of the standard ceteris paribus interpretation of estimated regression coefficients.

Relative to the treatment effects literature, which assumes that the treated units respond passively to treatment, the settable system framework permits the units to be active optimizing agents and explicitly permits these agents to interact with one another. A significant consequence is that agents may be affected not only by treatments applied to them specifically, but also by treatments applied to other agents. We thus explicitly remove the Stable Unit Treatment Value Assumption (SUTVA) standard in the treatment effects literature. The settable system framework permits precise definitions of interventions and counterfactuals and clarifies their relation. It further provides significant insight into the specification of the covariates required to ensure identification of causal effects of interest and into the central role of economic theory in specifying covariates. This, in turn, provides the basis for straightforward new tests of identification of causal effects, taken up in White and Chalak (2006).

Much of the treatment effects literature analyzes effects of single binary treatments. The settable systems analyzed here provide a framework for handling multiple treatments, each of which may be binary, categorical, or continuous. We thus provide causal foundations and identification results complementary to the treatment effects literature, including those that analyze more general treatments of this sort (e.g., Robins, Hernan, and Brumback, 2000; Lechner, 2001; Imbens and Newey, 2003; Hirano and Imbens, 2004; Robins, Hernan, and Siebert, 2004).

Relative to the artificial intelligence literature, the settable system framework provides a formal framework in which notions of mutual causality, central to economic concerns with optimization and equilibrium, are well defined and non-paradoxical. The settable system framework provides an underlying foundation from which standard directed acyclic graph-based causal structures can emerge as special cases, without having to impose useful structure (e.g., acyclicality) as axiomatic. The settable system framework further accommodates both observable and unobservable attributes that may act as response and effect modifiers.

Section 2 gives formal definitions of the settable systems supporting our further analysis. We blend ideas explicit and implicit in the classical structural equations framework with concepts from the treatment effect literature and the DAG approach of the artificial intelligence literature. The result is a framework possessing two complementary components, one providing explicit stochastic structure and the other providing formal causal structure.

As our causal structure accommodates definitions of causal effects for simple (binary), categorical, and continuous causes, we consider two kinds of effects: marginal ceteris paribus effects, involving an infinitesimal change to a continuous cause of interest holding all other causes of interest fixed; and effects of more general interventions, involving non-infinitesimal changes to multiple causes of interest. Because the variables involved in the responses underlying these effects typically cannot all be observed and cannot therefore be expliclty controlled, we study expectations or averages of these effects, conditioned on observable covariates.

In Section 3 we begin our study of "structural identification," by which we mean the equality of certain covariate-conditioned counterfactual averages with stochastically meaningingful and empirically accessible standard conditional expectations. In contrast, "stochastic identification," which is the focus of the study of identification in the classical structural equations framework, is concerned with issues involving observational equivalence. For example, perfect multicollinearity in a linear regression constitutes a failure of stochastic identification. Stochastic identification is neither necessary nor sufficient for structural identification.

Section 3 first treats covariate-conditioned average effects of general interventions and covariate-conditioned average marginal effects, extending the conditional average treatment effects of binary interventions studied by Abadie and Imbens (2002). As Imbens (2004) notes, consideration of such averages permits analysis of the effects of interventions on any aspect of the conditional distribution of the response that can be expressed in terms of conditional moments. We discuss these conditional moment effects, as well as effects defined more generally in terms of optimizers of covariate-conditioned objective functions or of implicit moments. The results of Section 3 treat structural identification in a framework exploiting randomization or, more generally, conditional randomization.

An important concept that emerges in Section 3 is that of conditional exogeneity, a generalization of strict exogeneity. We show how conditional randomization can ensure conditional exogeneity, which in turn ensures standard unconfoundedness conditions of the treatment effects literature. (Rosenbaum and Rubin, 1983; Hirano and Imbens, 2004).

In Section 4 we study conditions for structural identification in the absence of randomization, based on covariates constructed using "predictive proxies," similar to those discussed by White (2005a). We provide formal treatment of general effects defined by covariate-conditioned moments or distributional aspects. An innovation here is the appearance of specific "discrepancy scores" and "effect discrepancy" measures that serve as formal quantifiers of the degree of departure from structural identification.

Section 5 discusses the relation between unconfoundedness and conditional exogeneity and the implications of this relation for contrasting the settable systems approach with the treatment effect and DAG frameworks. Section 6 contains a summary and concluding remarks, including directions for further research. Proofs of formal results are gathered into the Mathematical Appendix.

2 A Causally Structured Data Generating Process

2.1 Attribute-Indexed Settable Systems

We begin by specifying a mathematical framework that supports formal notions of cause and effect, a version of White's (2005b) settable systems.

Definition 2.1 Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, and let $\mathcal{A}$ be a non-empty Borel set. For *agents* $h = 1, 2, \ldots$, let *attributes* a_h belong to $\mathcal{A}$, and put $a \equiv \{a_h\}$. For $h = 0, 1, \ldots$, and $j = 1, 2, \ldots$, let *settings* $Z_{h,j} : \Omega \to \mathcal{R}$ be measurable functions. For $h = 1, 2, \ldots$ and $j = 1, 2, \ldots$, let *attribute-indexed response functions* $r_{h,j} : \mathcal{R}^\infty \times \mathcal{A}^\infty \to \mathcal{R}$ be measurable functions.

For $h = 1, 2, \ldots$ and $j = 1, 2, \ldots$, let $Z_{(h,j)}$ be the vector including every setting except $Z_{h,j}$, and define the *attribute-indexed settable variables* $\mathcal{X}_{h,j} : \{0, 1\} \times \Omega \to \mathcal{R}$ such that

$$\begin{aligned} \mathcal{X}_{h,j}(1, \cdot) &= Z_{h,j} \\ \mathcal{X}_{h,j}(0, \cdot) &= r_{h,j}(Z_{(h,j)}, a). \end{aligned}$$

For $h = 0$ and $j = 1, 2, \ldots$, let $\mathcal{X}_{h,j}\ (0, \cdot\ ) = \mathcal{X}_{h,j}\ (1, \cdot\ ) = Z_{h,j}$.

Put $Z \equiv \{\ Z_{h,j} : j = 1, 2, \ldots;\ h = 0, 1, \ldots\}$, $r \equiv \{r_{h,j}, j = 1, 2, \ldots;\ h = 1, 2, \ldots\}$, and $\mathcal{X} \equiv \{\mathcal{X}_{h,j}$, $j = 1, 2, \ldots;\ h = 0, 1, \ldots\}$.

The pair $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P),\ (\mathcal{A}, a, Z, r, \mathcal{X})\}$ is an *attribute-indexed settable system.* ■

This definition provides a framework in which can be represented the responses of each of a collection of *agents* ($h = 1, 2, \ldots$) to an environment consisting of the other agents in the collection. The notion of "agent" should be interpreted broadly. The agent may be an individual economic agent, such as a consumer or firm; a collection of economic agents, such as an industry or market; or a physical, natural, or technological process, such as a production function. This framework is thus suitable for analysis of systems in the social sciences as well as in the clinical or natural sciences.

Our focus is on economic systems. In non-economic contexts, our agents might instead be viewed as passive responders and called "units," "subjects," "cases," or "patients." We do not rule out the possibility of passive responders, but given the active and interactive nature of the components of economic systems, it is important for our terminology to reflect this. Moreover, the above interactive structure maps directly to the primitive structures on which economic theory operates.

Associated with each agent h is an attribute vector a_h belonging to the *attribute space* $\mathcal{A}$, a multi-dimensional space, with dimensions for each possible characteristic of the agent, including what kind of agent it is (individual, firm, market, technological process).

Attributes a_h are fixed characteristics of the agent, such as the race or gender of an individual. To the extent that a given attribute is not strictly immutable in some larger context (e.g., gender), we view it as determined in some more comprehensive system that nests the present system. Thus, within a given settable system, attributes can vary across agents but can never vary for a given agent.

The jth *response* of agent h, conveniently denoted $Y_{h,j} = \mathcal{X}_{h,j}(0, \cdot\ )$, is represented as depending on settings $Z_{(h,j)} = \mathcal{X}_{(h,j)}\ (1, \cdot\ )$ and agent attributes a. The key idea is that the response $Y_{h,j}$ is generated by setting all other variables of the system to $Z_{(h,j)}$, leaving the jth variable for agent h free to respond in whatever manner the agent may determine, represented as $r_{h,j}(Z_{(h,j)},\ a\ )$. The response will generally depend on the own attributes a_h of the agent, but is also allowed to depend on the characteristics of the other agents.

For example, $Y_{1,1}$ may represent the wage earned by an individual as it depends, among other things, on settings of the individual's education and ability, given that individual's race and gender attributes. Or $Y_{1,1}$ may represent the market price of a call option as it depends on the settings of the risk-free interest rate and underlying asset price, given that option's attributes of underlying asset, strike price, and expiration date. As another example, $Y_{1,1}$ may represent the optimal price of firm 1's product to various classes of buyers designated by buyer attributes (e.g., age, when implementing a senior citizen discount), when all other factors relevant for price determination, including prices set by other firms in the industry, cost and demand factors, etc., assume the value $Z_{(1,1)}$.

Although settings and attributes are conceptually distinct, it is convenient to have a way to refer to them jointly. As both play a fundamental role in determining responses, we will refer jointly to settings and attributes as *explanatory variables.* As Hoover (2001, pp. 147-148) notes, Wold argued for the use of the "explanatory variables" nomenclature in order to avoid referring directly to "causal" variables, given doubts then raised as to the meaning of causality. (Hoover (2001, section 7.1) provides enlightening discussion of the historical context.) Our use is distinct, in that within our circumscribed framework causal meanings will be crisply defined. Nevertheless, the spirit of the meaning is similar, in that explanatory variables are not necessarily causal: as discussed below, attributes are not causal.

The response equation $Y_{h,,j} = r_{hj}(Z_{(h,j)},\ a)$ corresponds directly to a structural equation in the classical structural equations framework. The response $Y_{h,j} = \mathcal{X}_{h,j}(0, \cdot\ )$ corresponds to the "left-hand side" or "dependent" variable, and the settings $Z_{(h,j)} = \mathcal{X}_{(h,j)}\ (1,\ .\ )$ and attributes a correspond to the "right-hand side" variables in this framework. A key difference between the classical structural equations and the response equations of the settable system is that classical structural equations are often "simultaneous." In contrast, the response equations by construction do not admit simultaneity. Simultaneity arises in classical structural equations when the same variables occur on both the left- and right-hand sides in a system of structural equations. In the settable system, simultaneity is ruled out, because when a settable variable appears on the left-hand side, it is as a response, $\mathcal{X}_{h,j}(0,\ \cdot\ )$, whereas when it appears on the right, it is as a setting, $\mathcal{X}_{h,j}(1,\ \cdot\ )$.

The distinction between settings and responses enforced by this structure can be viewed as a formalization of the device introduced in Strotz and Wold (1960) and later used by Fisher (1970) of "wiping out" equations of the system that otherwise govern the behavior of a given variable, and replacing these with arbitrary values ("settings" in the nomenclature here) of that variable. This device is explicitly at work in Pearl's DAG approach (e.g., Pearl, 1995, 1998) where it is referred to as "intervention." We use the term "setting" here to convey the same notion, reserving "intervention" for formal definition later in a closely related context. The response equation corresponds to Pearl's "functional causal equation" (see e.g., Pearl, 2000, p. 27). Here, however, we do not (yet) distinguish between observable and unobservable causes and we permit the appearance of attributes.

Our discussion above refers to agents $h = 1, 2, \ldots$; the definition also references settings $Z_{h,j}$ for $h = 0$. Nevertheless, $h = 0$ does not refer to an agent. Rather, $Z_0 \equiv \{Z_{0,j} : j = 1, 2, \ldots\}$ refers to a particular collection of settings that have a unique function in the settable system, playing one of several different important roles. For example, certain agents may have responses that do not depend on any other variable of the system. For such cases, we may think of Z_{0,j^*} for a given index j^* as specifying the "default" value for such a response, so that for some indexes h and j we have $Y_{h,j} = \mathcal{X}_{h,j}(0, \cdot) = Z_{0j^*}$ for all values of the other settings. A second role is that of "initial" values, so that the impact of the other variables is to change the response $Y_{h,j} = \mathcal{X}_{h,j}(0, \cdot\ )$ to something different than its initial value, Z_{0,j^*}, which appears as an argument of its response function.

A third role for the settings Z_0 is as "fundamental" variables, in that the settable variables $\mathcal{X}_0$ have settings $Z_0 = \mathcal{X}_0(1, \cdot)$ that may impact other variables in the system, but their responses $\mathcal{X}_0(0, \cdot)$ do not depend on any other variables of the system. This contains the default and initial value roles as special cases. This unique property motivates our imposition of the convention $\mathcal{X}_0(0, \cdot) = \mathcal{X}_0(1, \cdot)$. The fundamental variables share certain features with the exogenous variables of the classical structural equations framework, but also differ in important ways. We discuss this further below.

We can relate the treatment effect framework of Rubin (1974) to Definition 2.1 using Holland's (1986) mathematical description of Rubin's framework. According to Holland, Rubin's framework can be represented as a quadruple (U, K, Y, S). Here, U and K are sets, Y is a mapping, $Y : U \times K \to \mathbb{R}$, and S is a mapping, $S : U \to K$ The set U contains the population of "units" whose responses to treatment are of interest. The set K contains the admissible values of treatment "labels" (e.g., $K = \{t, c\}$, with "t" for treated, "c" for control), and Y is the response function, such that $Y(u, t)$ is the response of unit u to treatment t. The function S is the treatment assignment function, so that unit u receives treatment $S(u) \in K$. Holland also references other measurements $X : U \times K \to \mathbb{R}$ (e.g., covariates), and a special type of measurement $X : U \to \mathcal{R}$ explicitly identified as attributes.

Holland's function $Y : U \times K \to \mathbb{R}$ corresponds to a response function, say $r_{\cdot,1} : \mathbb{R}^\infty \times \mathcal{A}^\infty \to \mathbb{R}$ in Definition 2.1. The units u belonging to Holland's U correspond to the agents $h \in \{1, 2, \ldots\}$ of Definition 2.1. Holland's K is the analog of $\mathbb{R}^\infty$ the domain of $r_{\cdot,1}$ By restricting $r_{\cdot,1}$ to depend only on a single variable (treatment) and not to depend on attributes (except through the agent index), we obtain Holland's mapping Y. Holland's treatment assignment $S(u)$ corresponds to a setting, say

$Z_{h,2}$, in Definition 2.1. Holland's covariates X correspond to other responses or settings (including fundamental settings) in Definition 2.1, and Holland's attributes $X(u)$ correspond to attributes a_h.

Thus, Holland's formulation of Rubin's treatment effect framework, including covariates and attributes, is contained in Definition 2.1. Definition 2.1 contains considerably more, however. The stochastic structure $(\Omega, \mathcal{F}, P)$ made explicit in the settable system framework is easily adjoined to Rubin's framework, so this is not a major difference. The most important difference is the possibility of agent optimizing behavior and interaction with other agents explicitly permitted by Definition 2.1.

The possibilities are extensive, but as a concrete example, Definition 2.1 provides a structure in which one can analyze trade. Consider an experiment in which agents differ by the attribute of preferences and possess varying initial endowments of several goods (fundamental settings). A subset is "treated" by receiving a standardized additional endowment; a "control" group does not receive this increment. Then the agents are permitted to trade, with the responses of interest being agents' consumption of each good after trading ceases. In such systems, the treatment of one agent generally affects the responses of others. In the treatment effect literature, this is typically ruled out by the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980, 1986): treatments affect only treated units. We remove this: the response $r_{h,j}$ may depend on the settings of all other variables $Z_{(h,j)}$, not just the settings for that agent.

Another difference between Rubin's framework and the present settable system framework is that the responses may differ not only on the basis of the agent's own attributes, but also on the basis of the attributes of the other agents that a given agent may interact with, either directly or indirectly. This permits price discrimination, among other possibilities.

2.2 Heuristics of Cause and Effect

The formal notion of response functions makes possible formal definitions of causality. It is helpful, however, to begin with some heuristics. The essential idea underlying the notion of "cause" here is that if, other things equal, the response of a given settable variable $\mathcal{X}_{h,j}$ does not depend on the setting of some other settable variable $\mathcal{X}_{i,k}$ ($i \neq h$ or $j \neq k$), then $\mathcal{X}_{i,k}$ does not cause $\mathcal{X}_{h,j}$. Otherwise, $\mathcal{X}_{i,k}$ causes $\mathcal{X}_{h,j}$. We give formal definitions below, but this heuristic understanding suffices for now. Thus we write $\mathcal{X}_{i,k} \not\Rightarrow_S \mathcal{X}_{h,j}$ if $\mathcal{X}_{i,k}$ does not cause $\mathcal{X}_{h,j}$ according to this heuristic definition, and we write $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$ if $\mathcal{X}_{i,k}$ causes $\mathcal{X}_{h,j}$. We explicitly reference the settable system $\mathcal{S}$, as causal relationships generally depend on the variables specifically referenced in the settable system.

We emphasize that causality is properly defined with respect to the *settable variables* $\{\ \mathcal{X}_{h,j}\ \}$, and not with respect simply to the random variables representing settings or responses or to events involving these random variables. This reflects the explicit distinction provided by the settable system $\mathcal{S}$ between the *stochastic structure* $(\Omega, \mathcal{F}, P)$ and the *causal structure* $(\mathcal{A},\ a,\ Z,\ r,\ \mathcal{X})$.

In accord with this understanding, we may also refer to settable variables as "causal variables." Causal variables $\mathcal{X}_{h,j}$ thus generate random variables $\mathcal{X}_{h,j}(1, \cdot\ )$ and $\mathcal{X}_{h,j}(0, \cdot\ )$.

Although the present framework is substantially inspired by the work of Pearl (1988, 1993a, 1993b, 1995, 1998, 2000) and of SGS, our focus on settable variables distinguishes it from these prior approaches. There, causality is defined either in terms of events or in terms of random variables, and it is formally *axiomatic* that if A causes B, then B does not cause A. (See for example SGS, p. 42.) It is this axiom that rules out the perplexities of "simultaneous causation" in Pearl's framework and that enforces the acyclicality of the directed graphs embodying the causal relations there.

Such an axiom is antithetical to economics, given the central and explicit interest of economics in the processes of optimization and equilibrium. For example, in Nash equilibrium, each economic agent responds in a mutually consistent way to the behavior of others. Each agent's behavior is related to that of others through the reaction function for that agent; formally, these are response functions, as defined above. In the sense just given, each agent's behavior is caused by that of the others. In the settable system framework, this is an example of the well-defined notion of "mutual causality."

Indeed, the inability to explicitly handle such cases may be one reason why the important advances of Pearl and his colleagues have had less impact in economics than elsewhere. (See, for example, Shipley (2000) for an extensive treatment of Pearl's framework from the point of view of biology.) In contrast, the present settable system framework easily accommodates mutual causality, without creating the perplexities of simultaneous causation.

To substantiate this claim, consider that causality, as defined heuristically above, admits four possibilities: $\mathcal{X}_{i,k} \Rightarrow\!\!|_S \mathcal{X}_{h,j}$ and $\mathcal{X}_{h,j} \Rightarrow\!\!|_S \mathcal{X}_{i,k}$ (mutual non-causality); $\mathcal{X}_{i,k} \Rightarrow\!\!|_S \mathcal{X}_{h,j}$ and $\mathcal{X}_{h,j} \Rightarrow_S \mathcal{X}_{i,k}$ (directed causality from $\mathcal{X}_{h,j}$ to $\mathcal{X}_{i,k}$); $\mathcal{X}_{h,j} \Rightarrow\!\!|_S \mathcal{X}_{i,k}$ and $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$ (directed causality from $\mathcal{X}_{i,k}$ to $\mathcal{X}_{h,j}$); and $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$ and $\mathcal{X}_{h,j} \Rightarrow_S \mathcal{X}_{i,k}$ (mutual causality). The latter case involves no paradoxes or perplexities, because it arises from two separate scenarios, one in which one variable is free to respond to the settings of the others, and another in which the other variable is free to respond to the settings of the remaining variables.

The settable system framework also admits causal "cycles." That is, we may have $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$,

$\mathcal{X}_{h,j} \Rightarrow_S \mathcal{X}_{g,l}$, and $\mathcal{X}_{g,l} \Rightarrow_S \mathcal{X}_{i,k}$. This can be viewed as a form of "indirect" mutual causality. Again, however, there is no paradox created in the settable system, due to the separation enforced between settings and responses. In contrast, the causal structures of Pearl's framework rule out such possibilities by working only with acyclic structures.

The settable system framework can be viewed as an extension of Pearl's framework. Dawid's (2002) "intervention directed acyclic graphs" (IDAGs) extend Pearl's framework to explicitly accommodate "interventions" (direct changes) to each variable of the system. The settable system framework constitutes an extension of Dawid's IDAGs, in that the settable system can be represented as a collection of modified IDAGS, one for each settable variable of the system, each representing the response (i.e. causal) dependencies for a given settable variable. Each such IDAG is modified in that it admits interventions only, and not to all variables of the system, but rather to all but the settable variable whose response it represents. This preserves many of the powerful conceptual advantages of Pearl's approach, but without assuming away the mutually causal relationships central to economics.

Notions of "effect" can be straightforwardly defined in terms of the response function. Specifically, when the derivative exists, the *marginal ceteris paribus effect on $\mathcal{X}_{h,j}$ of $\mathcal{X}_{i,k}$ given $z_{(h,j)}$* is $(\partial r_{h,j}/\partial z_{i,k})(z_{(h,j)}, a)$. The "reciprocal" marginal ceteris paribus effect on $\mathcal{X}_{i,k}$ of $\mathcal{X}_{h,j}$ *given* $z_{(i,k)}$ is $(\partial r_{i,k}/\partial z_{h,j})(z_{(i,k)}, a)$. There is no necessary inverse relation between reciprocal effects.

When $\mathcal{X}_{i,k}$ does not cause $\mathcal{X}_{h,j}$ its marginal effects are zero everywhere. Nevertheless, the marginal effect of $\mathcal{X}_{i,k}$ on $\mathcal{X}_{h,j}$ can be zero for certain values of $z_{(h,j)}$, even when $\mathcal{X}_{i,k}$ does cause $\mathcal{X}_{h,j}$.

Effects involving several settings can also be readily defined. For example, the *effect on $\mathcal{X}_{h,j}$ of the intervention $z_{(h,j)} \to z^*_{(h,j)}$* to $\mathcal{X}_{(h,j)}$, is the difference

$$\Delta r_{h,j}(z_{(h,j)}, z^*_{(h,j)}, a) \equiv r_{h,j}(z^*_{(h,j)}, a) - r_{h,j}(z_{(h,j)}, a).$$

In defining this effect, we make use of the notion of an intervention, as is standard in the literature. At one level, this is purely formal: an effect is simply a property of the response function involving evaluation at two different values of its first argument. By referring to an intervention $z_{(h,j)} \to z^*_{(h,j)}$, we specify precisely the values we mean.

At another level, however, the word "intervention" implies an active manipulation of some kind, suggesting that an intervention can be thought of as a change in the settings of $\mathcal{X}_{(h,j)}$ from $z_{(h,j)}$ to $z^*_{(h,j)}$. It must be clearly understood, however, that the manipulation involved in an intervention is a conceptual one and not an empirically observable operation. What is being compared in the definition of effect is the response under two possible outcomes of the settings, $z^*_{(h,j)} = Z_{(h,j)}(\omega^*)$ and $z_{(h,j)} = Z_{(h,j)}(\omega)$ for ω^*, $\omega \in \Omega$. Accordingly, we formally define an *intervention* as a pair (ω, ω^*) of elements of Ω. Any effect of interest can then be defined using this formalism.

Empirically, only one realization from Ω operates to generate the observed data; for a given intervention (ω, ω^*), this realization need not correspond to either ω^* or ω . Put differently, if we view the realization from Ω generating the data as "fact," then for a given intervention, either ω^* or ω or *both* must be counterfactual. For convenience and consistent with common usage, we will speak here of interventions as "changes in settings," as suggested above, but with the conceptual and counterfactual nature of the manipulations involved clearly in mind.

In contrast to settings, attributes do not depend on ω and are *fixed* in this precise sense. Within a given settable system, attributes are therefore not subject to even the conceptual manipulations applicable to settings. We thus adopt the convention that attributes do not have causal properties, consonant with Holland and Rubin's dictum, "No causation without manipulation" (Holland, 1986, p. 959). Although previously implicit, this makes it fully explicit that attributes *cannot act as causes and cannot have effects.* Rather, their role, when present, is to modify responses and effects.

By formalizing notions of interventions and the effects that arise from them as we have, we ensure that these concepts are defined entirely *within* the settable system. Recall that interventions allow us to manipulate realizations of settings of interest. Another way of manipulating realizations is to change the mappings $Z_{h,j}$ defining the settings, while keeping the element ω fixed. Such alterations constitute what we shall call *modifications to the causal structure* and are therefore *modifications of the settable system.* Modifications have a significant role to play in analyzing causality, but they are not necessary for defining concepts of effect, so we do not employ them for this purpose.

Observe the distinction between counterfactuals, which involve realizations from Ω other than those generating the data and which thus involve operations *within* the settable system, and modifications, which involve operations *on* the settable system.

Modifications to the settable system can play several important roles. Specifically, certain modifications correspond to the exercise of experimental control. For example, modifying the *stochastic* structure by modifying the underlying probability measure P can result in different patterns of treatment assignment. As is well known, randomized treatment assignments have particular utility in the attempt to learn about causal effects, so exercising experimental control to enforce a P implementing randomization will generally be advantageous.

Modifying the *causal* structure can also be viewed as an exercise of experimental control. For example, the choice of the mappings Z can determine whether treatments are binary, categorical, or continuous. These choices can all be viewed as part of the design of the experiment.

Modification of the attribute space and the attribute sequence permits a choice of the population of agents studied: by this means one may focus a given study on women, on firms in a given industry, or on electronic circuits.

Accordingly, one may view an attribute-indexed settable system as an *experiment* and think of certain modifications to the settable system as corresponding to the exercise of experimental control implementing experimental design.

Not all modifications of the settable system require this degree of experimental control, however. An especially important modification involves only conceptual manipulations. These impact the response functions by partitioning the responses of the settable system in useful ways. As we discuss below, partitions that lead to recursive structures can be particularly advantageous.

Although the settable system framework admits mutual causality without paradox, systems in which mutual causality is absent are simpler to analyze. Just as classical structural equations can be manipulated to yield reduced forms in which simultaneity is absent, settable systems provide a formal structure that can be manipulated to yield further settable systems in which mutual causality is absent. To keep the analysis here to a manageable level, we focus here on identifying causal effects in such systems. Elsewhere we may tackle the issues involved with identifying causal effects in systems admitting mutual causality.

2.3 Partitioned Settable Systems and Formal Causality

To resolve mutual causalities, we begin by considering systems in which the settable variables are partitioned into blocks, such that the responses of variables in a given block are jointly determined by settings of variables outside the given block, but not by settings of other variables inside that block. We formalize this partitioning as follows.

Definition 2.2 Let $(\Omega, \mathcal{F}, P)$, $\mathcal{A}$, a, and Z be as in Definition 2.1. Suppose that settings of $\mathcal{X}_{h,j}$ are given by $\mathcal{X}_{h,j}\ (1, \cdot\ ) = Z_{hj}$, $j = 1, 2, \ldots$, $h = 0, 1, \ldots$, and let $\Pi = \{\Pi_b\}$ be a partition of the ordered pairs $\{(h, j)$: $j = 1, 2,\ldots$; $h = 1, 2, \ldots \}$. Suppose there exists a countable sequence of measurable functions $r^{\Pi} \equiv \{r^{\Pi}_{h,j}\}$ such that for all (h, j) in Π_b the responses $Y_{h,j} = \mathcal{X}_{h,j}\ (0, \cdot\ )$ are jointly determined as

$$Y_{h,j} = r^{\Pi}_{h,j}(Z_{(b)}, a), \qquad b = 1, 2, \ldots,$$

where $Z_{(b)}$ is the countable vector containing $Z_{i,k}$, $(i,k) \notin \Pi_b$. Then $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P),\ (\mathcal{A}, a,\ Z, \Pi, r^{\Pi}, \mathcal{X})\}$ is an *attribute-indexed partitioned settable system.* ■

The response functions of Definition 2.1 correspond to the elementary partition in which the typical block is $\Pi_b = \{(h,j)\}$. Other natural partitions are those that group together all responses governed by a given agent or collection of agents. In the former case, this yields $\Pi_b = \{(b,j),\ j = 1, 2, \ldots\}$ (here $b = h$), permitting the responses to be viewed, for example, as the outcome of an agent's (joint) optimization problem. An example of the latter case arises in studying the price behavior of a cartel, in which case the partition may act to group the price responses of all cartel members.

The partition has significant implications for the interpretation of the response functions. The response functions $r^{\Pi}_{h,j}$ are now *joint response functions.* They describe the joint response of settable variables within a block to settings of variables *outside* that block. This eliminates mutual causality *within* the block because by construction, the response of any variable in a given block cannot depend on the settings of other variables in the same block.

For example, consider a firm optimizing its factor inputs of capital and labor. The joint response functions for capital and labor specify optimal choices for capital and labor in response to rents, wages, and other relevant factors. The response functions of the unpartitioned system embody short run responses separately describing (1) the firm's optimal choice of labor given settings of capital, wages, and rents, and (2) the firm's optimal choice of capital given settings of labor, wages, and rents. Mutual causalities exist between the capital and labor responses for the unpartitioned system, but these are absent within the block for the partitioned system.

Although the partition just discussed removes mutual causalities within a block, it still permits mutual causalities between blocks or causal cycles. To make this precise and to define structures in which both direct and indirect mutual causality are fully absent, we need a formal definition of causality. We now have a sufficiently rich formal framework in which to state the required definition. We make explicit one more item of notation. Specifically, for a given block Π_b of a given partition Π, we write $z_{(b),(i,k)}$ to refer to the vector containing elements corresponding to every setting of the system except for those indexed by the elements of Π_b and by $(i,k) \notin \Pi_b$.

Definition 2.3 Let $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P),\ (\mathcal{A}, a, Z, \Pi, r^{\Pi}, \mathcal{X})\}$ be an attribute-indexed partitioned settable system. For given positive integer b, let $(h,j) \in \Pi_b$. (i) If for given $(i,k) \notin \Pi_b$ the function $z_{i,k} \to r^{\Pi}_{h,j}(z_{(b)}, a)$ is constant in $z_{i,k}$ for every $z_{(b),(i,k)}$, then we say $\mathcal{X}_{i,k}$ *does not cause* $\mathcal{X}_{h,j}$ *in* $\mathcal{S}$ and write $\mathcal{X}_{i,k} \Rightarrow\!\!|_S \mathcal{X}_{h,j}$. Otherwise, we say $\mathcal{X}_{i,k}$ *causes* $\mathcal{X}_{h,j}$ in $\mathcal{S}$ and write $\mathcal{X}_{i,k} \Rightarrow_S \mathcal{X}_{h,j}$. (ii) For $(i,k),(h,j) \in \Pi_b$, $\mathcal{X}_{i,k} \Rightarrow\!\!|_S \mathcal{X}_{h,j}$. ■

In defining non-causality, this definition formalizes the heuristic given above that the response of $\mathcal{X}_{h,j}$ does not depend on the setting of $\mathcal{X}_{i,k}$. For simplicity, we permit the settings to take any real value. More refined definitions arise in the obvious way by restricting the settings to subsets of $\mathbb{R}$ suitable to the character of the particular causal variable (e.g., binary, categorical, bounded continuous). In these refinements, only the admissible values of $z_{i,k}$ and $z_{(b),(i,k)}$ are relevant. Extensions of this definition are also straightforward. Specifically, instead of simply being real-valued, the settable variables may take values that are elements of a function space, such as a conditional probability density. Among other things, this accommodates a causal role for beliefs.

The notion of cause defined here corresponds to the notion of *direct* cause in Pearl's DAG framework. We do not make a distinction here between direct and indirect causes, as we prefer to emphasize the role of the settable system $\mathcal{S}$, and in particular the roles of the specified variables $\mathcal{X}$ and/or the partition Π, in defining causal relations. Causes that would be viewed as indirect in the DAG framework become direct causes under a suitable re-partition or respecification of the variables of the settable system. Only a single causal concept is therefore required.

By construction, dependence of response in a partitioned system is possible only between settable variables in different blocks. For completeness, part (ii) of the definition explicitly states the necessary absence of causality arising from this construction. This includes the convention that the causality relation is non-reflexive, as $\mathcal{X}_{i,k} \Rightarrow\!\!|_S \mathcal{X}_{i,k}$: a settable variable does not cause itself. Nevertheless, for convenience in using the notation $\mathcal{X}_{i,k} \Rightarrow\!\!|_S \mathcal{X}_{h,j}$, it will be implicitly understood that the referenced variables belong to different blocks, unless it is explicitly stated to the contrary.

The meaning of the notations $\Rightarrow\!\!|_S$ and $\Rightarrow_S$ extends to accommodate disjoint sets of multiple settable variables appearing on the left and right by the convention that the indicated relation holds pairwise between each possible pair of left-hand and right-hand settable variables. Thus, $\mathcal{X}_i \Rightarrow\!\!|_S \mathcal{X}_h$ means that for all k and j, $\mathcal{X}_{i,k} \Rightarrow\!\!|_S \mathcal{X}_{h,j}$.

2.4 Recursive and Reduced Settable Systems

In the standard DAG framework, causal relations are necessarily anti-symmetric, as it is an axiom that "if A causes B, then B does not cause A" (e.g., SGS, p. 42). Our structure so far does not require this. We now define a particular settable system in which anti-symmetry finally appears.

Definition 2.4 Let $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P), (\mathcal{A}, a, Z, \Pi, r^{\Pi}, \mathcal{X})\}$ be an attribute-indexed partitioned settable system. For $b = 1, 2, \ldots$, let $\mathcal{X}_{[b]}$ denote the vector containing the settable variables $\mathcal{X}_{h,j}$ for $(h, j) \in \Pi_b$ and let $\mathcal{X}_{[0]} = \mathcal{X}_0$. If Π is such that $\mathcal{X}_{[b]} \not\Rightarrow_S \mathcal{X}_{[0]}, \ldots, \mathcal{X}_{[b-1]}$, $b = 1, 2, \ldots$, then S is an *attribute-indexed recursive settable system.* ■

Definition 2.4 imposes a specific block recursive structure on the partition that eliminates mutual causalities and cycles. This is an analog of the classical recursive systems analyzed by Simon (1953), Wold (1954, 1956) and Fisher (1961, 1966), among others. In the structural equations framework, recursivity eliminates simultaneities. The motivation here is different, however, as the settable system *already* eliminates simultaneity. Instead, our goal is a simpler analysis of causal effects.

The analogy with classical recursive structures is loose, at best. Specifically, in the classical block recursive systems, the responses in a given block may be jointly determined, (e.g., Fisher, 1966, pp 99-100). Here, however, the responses in a given block are explicitly *not* mutually causal; the blocks here thus correspond to the *reduced forms* of the classical (simultaneous) blocks. Further, the unobservable disturbances in the classical case obey a block diagonal covariance structure. Here, this need not hold, because unobservables in a given block may have effects on each succeeding block.

The ordering of the blocks permits us to speak meaningfully about "levels" of the partition, and about "successors" (higher level blocks or their elements) and "predecessors" (lower level blocks or their elements). At any given block level b, the recursive system groups responses so as to ensure that whenever $\mathcal{X}_{i,k}$ causes $\mathcal{X}_{h,j}$, we do not have $\mathcal{X}_{h,j}$ causing $\mathcal{X}_{i,k}$. That is, successors do not cause predecessors. Mutual causality and cycles are now absent, embodying the anti-symmetry of the DAG framework. Here anti-symmetry is a restriction on a more general structure, rather than an axiom.

So far, we have not explicitly considered the role of time in our framework, as there has been no formal need for it. Nevertheless, time can play an important role by imposing natural recursive structure. To specify this role, we first recognize that settings and responses may be indexed by a special identifier, "time," such that settings or responses for different time values are necessarily conceptually distinct. For a given agent h, this means that the indexes j distinguishing relevant settings or responses implicitly depend on time, such that settings or responses for different times, say $t \neq t'$, necessarily have $j \neq j'$. Making this dependence explicit for the moment, we consider settable variables indexed by (h, j, t), where (j, t) now plays the role previously played by j alone.

With time explicit, the settable system obeys *weak time structure* if for all (i, k) and (j, h) $\mathcal{X}_{i,k,t}$ $\Rightarrow\!|_S$ $\mathcal{X}_{h,j,\tau}$ for all $t > \tau$, $t = 1, 2, \ldots$. It obeys *strong time structure* if for all (i, k) and (j, h) $\mathcal{X}_{i,kt} \Rightarrow\!|_S \mathcal{X}_{h,j,\tau}$ for all $t \geq \tau$, $t = 1, 2, \ldots$. Strong time structure thus rules out "contemporaneous" causality. In the classical structural equations framework, this eliminates simultaneity (cf. Wold, 1954, 1956; Cartwright, 1989). Here this eliminates mutual causality. Weak time structure permits simultaneity or mutual causality.

Clearly, either time structure contains recursive elements. Strong time structure apparently suffices to fully order the system; on the other hand, weak time structure delivers only a partial ordering. In practice, empirical researchers tend to work with weak time structures. Economic theory then plays a key role in completing the ordering of the recursive structure, specifying which causal variables may or may not cause which other causal variables.

It may at first seem counterintuitive that weak time structure is preferred in practice to strong time structure, as contemporaneous causation may seem both physically and metaphysically unappealing. This preference may be understood, however, as a natural response to the fact that in practice the *operational* time scale at which responses actually occur or evolve may be much smaller than the *observational* time scale on which responses can be measured. As the researcher can only work with the measured responses, contemporaneous causation may not afford too bad an approximation to the observed phenomena. This then necessitates careful consideration of mutual causality and the further recursive relations in addition to time that remove it as a complicating element.

Although our present settable system framework thus permits explicit consideration of time and dynamics, to maintain a manageable scope for the present analysis, we revert to our previous implicit treatment, recognizing nevertheless that the recursive structures treated henceforth may arise in significant part from temporal relationships.

Recursive settable systems are especially convenient because, as we now show, the settings of predecessor variables driving a given response can be viewed as responses to further predecessor settings. Nor is it necessary to be fully explicit about such predecessor responses. This is especially useful in economics, where both the ability to directly set potential causes of interest and the full understanding of their genesis otherwise may be absent.

To justify this claim, we introduce some convenient notation. Let $r^{\Pi}_{[b]}$ denote the joint response functions for levels $b = 1, 2, \ldots$ of a recursive settable system such that

$$\mathcal{X}_{[b]}(0, \cdot) = r^{\Pi}_{[b]}(\mathcal{X}_{[0]}(1, \cdot), \ldots, \mathcal{X}_{[b-1]}(1, \cdot)).$$

For simplicity, we leave attributes implicit. Next, we introduce the notion of *canonical settings.* Let $Z_{[b]} = \mathcal{X}_{[b]}(1, \cdot)$ denote the settings of $\mathcal{X}_{[b]}$, $b = 0, 1, \ldots$. We define canonical settings recursively as

$$Z_{[b]} = r^{\Pi}_{[b]}(\mathcal{X}_{[0]}(1, \cdot), \ldots, \mathcal{X}_{[b-1]}(1, \cdot)), \qquad b = 1, 2, \ldots.$$

This generates valid settings, because for each b, $Z_{[b]}$ is a measurable function of settings, which are themselves measurable functions. These settings are canonical in the sense that by construction $Z_{[b]} = \mathcal{X}_{[b]}(0, \cdot)$, $b = 0, 1, \ldots$, so that

$$Z_{[b]} = r^{p}_{[b]}(\mathcal{X}_{[0]}(0, \cdot), \mathcal{X}_{[1]}(0, \cdot), \ldots, \mathcal{X}_{[b-1]}(0, \cdot)).$$

Consequently, *these are the settings generated in the absence of experimental control.*

When the settings are canonical, we can view the settings driving the response at any level as responses to predecessor settings. Put more strongly, with canonical settings in a recursive settable system, the distinction between settings and responses vanishes. We formalize this as follows.

Definition 2.5 Let $\mathcal{S} \equiv \{(\Omega, \mathcal{F}, P), (\mathcal{A}, a, Z, \Pi, r^{\Pi}, \mathcal{X})\}$ be an attribute-indexed recursive settable system. Suppose the settings are the *canonical settings* such that

$$\mathcal{X}_{[b]}(1, \cdot) = Z_{[b]} = r^{\Pi}_{[b]}(\mathcal{X}_{[0]}(1, \cdot), \ldots, \mathcal{X}_{[b-1]}(1, \cdot)), \qquad b = 1, 2, \ldots$$

Then $\mathcal{S}$ is an *attribute-indexed canonical recursive settable system.* ■

In the classical structural equations framework, recursivity eliminates simultaneity, but it does not by itself eliminate "endogeneity," that is, the possible correlation of observable right-hand side variables with unobservable causes. This correlation arises from correlations between the unobservable causes driving responses in different levels of the system, and it renders standard estimation methods such as least squares inconsistent when applied directly to the structural equation of interest.

In the classical framework, the reduced form for given dependent variables eliminates endogeneity as well as simultaneity. The reduced form is obtained by solving a system of structural equations to express the "endogenous" (dependent) variables of interest as functions of observable exogenous variables, independent of unobservable "error terms." In the recursive case, simple substitution suffices. In the simultaneous case, the reduced form equations must represent the fixed point (assuming it exists) of the underlying simultaneous equations. That is, the classical reduced form represents an "equilibrium" relationship among the simultaneous equations (see also Matzkin, 2005).

The analog of the reduced form in the settable system framework provides a useful contrast to the classical reduced form. For any canonical recursive settable system, substitution gives

$$\mathcal{X}_{[1]}(0,\cdot) = r^{\Pi}_{[1]}(\mathcal{X}_{[0]}(1,\cdot))$$

$$\mathcal{X}_{[2]}(0,\cdot) = r^{\Pi}_{[2]}(\mathcal{X}_{[0]}(1,\cdot), r^{\Pi}_{[1]}\{\mathcal{X}_{[0]}(1,\cdot)\}),$$

$$\mathcal{X}_{[3]}(0,\cdot) = r^{\Pi}_{[3]}(\mathcal{X}_{[0]}(1,\cdot), r^{\Pi}_{[1]}(\mathcal{X}_{[0]}(1,\cdot)), r^{\Pi}_{[2]}\{\mathcal{X}_{[0]}(1,\cdot), r^{\Pi}_{[1]}[\mathcal{X}_{[0]}(1,\cdot)]\}),$$

and so forth, expressing each response in terms of the fundamental variables $\mathcal{X}_{[0]}$, so that

$$\mathcal{X}_{[b]}(0,\cdot) = r^{\Pi^0}_{[b]}(\mathcal{X}_{[0]}(1,\cdot)), \qquad b = 1, 2, \ldots,$$

say, where $r^{\Pi^0}_{[b]}$, denotes the "reduced response function" for the given system. This is the analog of the classical reduced form, representing the joint response function for the two-block partition Π^0 that groups the fundamental variables in one block and the remaining variables in the other.

The fundamental variables $\mathcal{X}_{[0]}(\ 1, \cdot\ )$ now appear analogous to the classical exogenous variables, but there are important differences. First, we have made no assumptions about observability: all, some, or none may be observable. Moreover, even if some are observable, we make no assumptions about independence between observables and unobservables. Nor will we assume the separability of the reduced response between its observable and unobservable arguments, as is common for the classical reduced form. The reduced response and the fundamental variables $\mathcal{X}_{[0]}(\ 1, \cdot\ )$ thus will not possess the standard properties assumed for the classical reduced form and its exogenous variables.

Moreover, our interest is on identifying and ultimately estimating the effects of variables in lower blocks on variables in higher blocks. As we shall shortly see, it is possible to identify these effects directly and easily without involving the reduced response. Accordingly, we do not consider it futher.

2.5 Constructing Recursive Settable Systems

It is not always obvious how to construct a recursive settable system from a given partitioned system when that system admits mutual causality or cycles; this is often true even in their absence. Fortunately, powerful methods for obtaining recursive settable systems are provided by graph theory.

Specifically, for finite settable systems (realistic in practice), proposition 1.4.3 of Bang-Jensen and Gutin (2001) (BJG) can be used to establish the existence of a recursive settable system whenever a partitioned settable system is acyclic. The DFSA algorithm of BJG (chapter 4.2) delivers the required ordering of the settable variables. Further results of BJG can be applied to systems with mutual causalities or cycles under mild conditions to ensure the existence of a partition of the system whose blocks possess an acyclic ordering. The SCA algorithm of BJG (chapter 4.4) delivers such a partition. The DFSA algorithm then delivers the ordering of the settable variables required for the recursive system. For brevity we leave further discussion of these methods aside here.

2.6 Settable Systems Generating Samples

In applications, interest often focuses on a modest number of comparable responses for a homogeneous class of agents, responding to settings with common meanings. For example, interest might focus on price responses of retail stores of a specific type at various locations to settings of wages, rents, transportation costs, density of nearby competitors, nearby consumer demographics, etc.

In such circumstances, it is convenient to identify a subset $\mathbf{A}$ of attribute space $\mathcal{A}$ designating the collection of comparable agents (firms, individuals, markets) whose responses are of interest and then to map the relevant comparable settable variables to a standardized collection of settable variables, $(\mathcal{Y}_h, \mathcal{Z}_h)$ say, permitting representation of the responses as

$$Y_h = r(Z_h, a_h, a_{(h)}) \qquad a_h \in \mathbf{A}, \qquad a_{(h)} \in \mathcal{A}_{(h)}.$$

In the applications considered here, we impose $\mathcal{Y} \not\Rightarrow_S \mathcal{Z}$, where $\mathcal{Y} \equiv \{\mathcal{Y}_h, a_h \in \mathbf{A}\}$ and $\mathcal{Z} \equiv \{\mathcal{Z}_h, a_h \in \mathbf{A}\}$, so we will be working at a particular level of a recursive settable system.

For simplicity and without essential loss of generality, we take Y_h and r to be scalar valued, so that Y_h is the response of $\mathcal{Y}_h$, that is $Y_h = \mathcal{Y}_h(0, \cdot)$. The potential causes $\mathcal{Z}_h$ may generate Z_h either as arbitrary settings, canonical settings, or a mix of these. We discuss this further below.

We split the dependence of the responses on the attributes into two parts, the "own attributes"

a_h of the agent, which by construction belong to $\mathbf{A}$, and the "other agent attributes" $a_{(h)}$, which represent the attributes of the agents that agent h intereacts with, such as customer characteristics when h is a firm. We let $\mathcal{A}_{(h)}$ designate the space in which $a_{(h)}$ takes its values.

So far, our discussion has focused on a finite or countable collection of agents, $h = 1, 2, \ldots$. This should be viewed as the population of agents relevant for the study of a given phenomenon. Usually, we do not observe the entire population, but rather a sample from the population. Here, sampling corresponds to generating a sequence of random positive integers say $\{\mathrm{H}_i\}$, governed by the underlying probability measure P such that realizations of a_h belong to $\mathbf{A}$. This yields sample responses

$$Y_{H_i} = r(Z_{H_i}, a_{H_i}, a_{(H_i)}), \qquad i = 1, 2, \ldots \ .$$

To simplify the notation, let $A_i \equiv (a_{H_i}, a_{(H_i)})$, and, engaging in a mild abuse of notation, write $Z_i = Z_{H_i}$ and $Y_i = Y_{H_i}$, so that sample responses can be represented as

$$Y_i = r(Z_i, A_i), \qquad i = 1, 2, \ldots .$$

For convenience in referring to this standard sampling situation in what follows, we say that the attribute-indexed recursive settable sytem $\mathcal{S}$ *generates a sample from* $\mathbf{A}$ *involving settable variables* $(\mathcal{Y}, \mathcal{Z})$. This enables us to avoid having to explicitly reference the totality of a given settable system in applications, while still implicitly referencing the full underlying structure.

Observe that although the dependence of agent responses on settings and attributes of other agents introduces stochastic dependence among the agents in the population, independent sampling results in independence of sample observations, conditional on the population values. Other forms of sampling may preseve certain aspects of the population dependence. These may be assessed in given applications and dealt with appropriately for purposes of estimation and inference using suitable laws of large numbers and central limit theorems.

3 Structural Identification with Randomization

Building on the foundations of Section 2, we now add structure enabling us to define and identify causal effects in the common situation in which not all relevant explanatory variables are observed.

3.1 A Causally Structured Data Generating Process

Our first assumption specifies causes of interest and designates which variables are observed and which are not. $\mathbb{N}$ denotes the natural numbers, including zero by convention.

Assumption A.1. Let an attribute-indexed recursive settable system $\mathcal{S}$ generate a sample from $\mathbf{A} \subset \mathcal{A}$ involving settable variables $(\mathcal{Y}, \mathcal{D}, \mathcal{Z})$ such that $\mathcal{Y} \not\Rightarrow_S (\mathcal{D}, \mathcal{Z})$. In addition:

(a) Let attributes $\{A_i \equiv (\tilde{A}_i, \ddot{A}_i)\}$ be a sequence of random vectors and let $(\mathcal{D}, \mathcal{Z})$ generate settings $\{(D_i, Z_i) \equiv (D_i, \tilde{Z}_i, \ddot{Z}_i)\}$ such that the joint distribution of $(D_i, \tilde{X}_i) \equiv (D_i, \tilde{Z}_i, \tilde{A}_i)$ is $\tilde{H}$ and the conditional distribution of $\ddot{X} = (\ddot{Z}_i, \ddot{A}_i)$ given $(D_i, \tilde{X}_i) = (d, x)$ is $\tilde{G}(\cdot \mid d, x)$ for all $i = 1, 2, \ldots$, where D_i is $\mathbb{R}^{k_1}$-valued, k_1 a positive integer, $\tilde{Z}_i$ is $\mathbb{R}^{k_2}$-valued, $k_2 \in \mathbb{N}$, $\ddot{Z}_i$ is $\mathbb{R}^{\infty}$-valued, $\tilde{A}_i$ is $\mathbb{R}^{\ell_1}$-valued, $\ell_1 \in \mathbb{N}$, and $\ddot{A}_i$ is $\mathbb{R}^{\infty}$-valued ;

(b) The responses $\{Y_i\}$ of $\mathcal{Y}$ are determined as

$$Y_i = r(D_i, Z_i, A_i), \qquad i = 1, 2, \ldots,$$

where r is an unknown measurable scalar-valued function;

(c) $\mathcal{D} \not\Rightarrow_S \mathcal{Z}$;

(d) The realizations of $Y_i, D_i, \tilde{Z}_i$ and $\tilde{A}_i$ are observed; those of $\ddot{Z}_i$ and $\ddot{A}_i$ are not. $\blacksquare$

This assumption explicitly focuses attention on the responses of the settable variable $\mathcal{Y}$ and whatever effects are embodied in the associated response function r, as modified by the relevant attributes. We consider a single response for simplicity and without essential loss of generality.

The recursive system imposes $\mathcal{Y} \not\Rightarrow_S (\mathcal{D}, \mathcal{Z})$; given A.1(b), we have that $\mathcal{Y}$ succeeds $\mathcal{D}$ and $\mathcal{Z}$. We view $\mathcal{D}$ and $\mathcal{Z}$ as potential causes of $\mathcal{Y}$, as we do not assume we know a priori which elements of $\mathcal{D}$ and $\mathcal{Z}$ are truly causal. By permitting $(\mathcal{D}, \mathcal{Z})$ to be countably dimensioned (A.1(a)), we ensure that we do not omit to reference any variable truly causal for $\mathcal{Y}$.

We reference $\mathcal{D}$ and $\mathcal{Z}$ separately to single out the finitely dimensioned vector $\mathcal{D}$ as containing the causal variables whose effects on $\mathcal{Y}$ are of primary interest. Interest in the effects of $\mathcal{Z}$ will be secondary. We thus refer to $\mathcal{D}$ as "causal variables of interest" and to the random variables D_i generated by $\mathcal{D}$ as "causes of interest." Recall, however, that causal relations properly hold between the underlying causal variables, rather than the random variables they generate. Given our secondary interest in the causal variables $\mathcal{Z}$, we refer to them as "ancillary causal variables" and to the random variables they generate, Z_i, as "ancillary causes." Rosenbaum (1984) calls these "concomitants." We adopt the present nomenclature to emphasize their causal role.

The causes of interest D_i correspond to "treatments" in the treatment effect literature. There, treatments are often assumed to be a binary scalar. Here, D_i may be a vector with binary, categorial or continuous components.

In A.1(a), the underlying probability measure P generates a sample of explanatory variables (D_i, Z_i, A_i), yielding the responses Y_i, $i = 1, 2, \ldots$ of A.1(b). We impose identical distribution for simplicity. This can be weakened substantially without changing the character of the results, but at the expense of considerable notational burden. We do not impose independence across observations, as this is not necessary for analyzing structural identification.

Assumption A.1(a) permits (D_i, Z_i) to be arbitrary settings, canonical settings, or a mix of the two: we may adopt whatever viewpoint is convenient or appropriate to the particular application. For example, in clinical settings, D_i may be a dosage set by the researcher. In non-experimental situations, D_i may be generated as a response to variables outside the researcher's control.

In A.1(a), we partition Z_i as $(\tilde{Z}_i, \ddot{Z}_i)$ and A_i as $(\tilde{A}_i, \ddot{A}_i)$. This accommodates the distinction between observed and unobserved explanatory variables, formalized in A.1(d). The conditional distribution $\tilde{G}$ explicitly permits dependence between observed and unobserved explanatory variables. This eliminates the possibility that the observed explanatory variables $(D_i, \tilde{X}_i)$ are exogenous in the classical structural equations sense (e.g., independent of the unobserved explanatory variables). Nevertheless, we shall see that identification of effects of interest is possible without classical exogeneity.

Drawing on the treatment effects literature, we refer to $\tilde{X}_i$, the observable explanatory variables that are not causes of interest, as *covariates*, although we do not require that these are measured prior to the setting of the causes of interest. Here the covariates consist of observable ancillary causes and attributes. Below we see how other variables can also serve as useful covariates.

Typically, economic theory does not provide strong guidance as to the functional form governing the response. Accordingly, A.1(b) imposes the mildest possible structure on r. Not only are the classical assumptions of linearity or separability between observables and unobservables not imposed, but neither are monotonicity, convexity, or any other particular but less specific structure. This provides a framework in which such restrictions can be tested. Alternatively, correctly imposing such structure may not only facilitate identification of certain features of interest not otherwise identifiable (e.g., as in Matzkin, 2003) but may also enhance estimation efficiency. To keep our analysis focused, however, we leave these possibilities aside here.

The response function r is assumed identical for all observations, but because the response depends explicitly on agent attributes, both observed and unobserved, this is without loss of generality. The unobserved heterogeneity admitted here corresponds to the "essential heterogeneity" of HV and HUV. When D_i is itself a response, essential heterogeneity is also permitted in determining D_i.

Note that one way of interpreting A.1(b) is that the response incorporates both observed and unobserved "fixed effects." The "fixed effects" are simply the own attributes component of the attributes, A_i. Given the preceding discussion, the "fixed effect" nomenclature appears singularly unhelpful. Attributes are indeed fixed, but they are not effects (i.e., changes in a response with respect to the setting of a potential cause) nor are they even potentially causal, given their immutability. They can, however, be usefully viewed as "fixed effect/response modifiers," that is, as effect and response modifiers that are fixed.

In A.1(c) we require $\mathcal{D} \Rightarrow\!\!|_S \mathcal{Z}$: $\mathcal{Z}$ may cause $\mathcal{D}$, but $\mathcal{D}$ does not cause $\mathcal{Z}$. Often, one can ensure this by measuring the setting of $\mathcal{Z}$ prior to assigning or measuring the setting of $\mathcal{D}$ (e.g., see Rosenbaum and Rubin, 1983). Nevertheless, this is not always convenient or appropriate; in such cases A.1(c) is essential to a full accounting of the effects of $\mathcal{D}$ on $\mathcal{Y}$, as in the absence of this requirement, one or more ancillary causes may mediate the effects of $\mathcal{D}$.

Rosenbaum (1984), Angrist and Krueger (1999), HV, and Wooldridge (2005) among others, discuss what happens when mediating causes are present, in violation of A.1(c). An important consequence of their presence is that any statistical analysis of such a system will fail to identify the indirect effects of $\mathcal{D}$, unless the statistical analysis explicitly accounts for the mediating responses.

The requirement that $\mathcal{D}$ not cause $\mathcal{Z}$ is a weak analog of Assumption A-6 of HUV and HV. Here we permit the possibility that elements of $\mathcal{Z}$ may cause $\mathcal{D}$ in a manner modified by attributes, both observed and unobserved. We explore this further below.

Thus, we focus on the total effect of the causes of interest, encompassing both direct and indirect effects. The response function (and any ensuing statistical analysis) must omit any variables that are themselves caused by the causes of interest. Because the A_i's are attributes, they cannot be causally affected by any settable variable of the system, and thus cannot mediate effects of $\mathcal{D}$.

Assumption A.1(d) formally specifies that we observe the responses generated by $\mathcal{Y}$. For simplicity and brevity, we do not permit censoring or truncation in observing the responses. Assumption A.1(d) further distinguishes $(D_i, \tilde{X}_i)$ as observable explanatory variables and $\ddot{X}_i$ as unobservable. To be observable in this sense, the explanatory variables must be measured *without error*. The requirement that k_1 be a positive integer ensures that we observe at least one cause of interest without error. Our methods here permit identification of causal effects only for accurately measured potential causes, so we treat potential causes that cannot be measured without error as necessarily ancillary. Other methods may permit identification of the effects of potential causes measured subject to error, but for simplicity we leave these aside.

In accord with this observability requirement, we view error-laden versions of otherwise relevant explanatory variables as structurally *irrelevant* for the response of interest. Thus, if no error-free version of a given relevant explanatory variable is available, we treat that variable as unobservable. It may therefore happen that $\tilde{X}_i$ is empty. In Section 4, we see how error-laden versions of structurally relevant explanatory variables can still play a useful role in identifying average effects of interest.

Assumption A.1(d) is not actually necessary for our structural identification results. Nevertheless, we state it to motivate the partition introduced in A.1(a).

3.2 Structural Identification of Covariate-Conditioned Average Effects

Because we generally do not observe realizations of all causes and attributes determining a given response, we cannot know or estimate r itself. This prevents us from knowing or estimating the effects defined in Section 2, even in the absence of the difficulties posed by their counterfactual nature.

Heckman (2005) forcefully argues that notions of “effect” are inherently counterfactual, and that such concepts have no meaning otherwise. As Dawid (2000) discusses, provided that the counterfactuals involved relate to certain properly behaved averages, they can be assigned empirical meaning. We now show how this works in the present framework under mild conditions, permitting us to define empirically meaningful counterfactual objects that can serve as a basis for empirically meaningful definitions of effects. Under further plausible conditions, these can be estimated.

First we give an elementary result describing the conditional expectation of the response given observable explanatory variables. We let $\text{supp}(D, \tilde{X})$ denote the support of $(D, \tilde{X})$. This is the smallest measurable set on which the joint density $d\tilde{H}$ of $(D, \tilde{X})$ integrates to one. Here and whenever convenient otherwise, we exploit the identical distribution assumptions (A.1(a, b)) to omit explicit reference to the subscript i. Corresponding to our notation for $\tilde{X}$ and $\ddot{X}$ we write $\tilde{x} \equiv (\tilde{z}, \tilde{a})$ and

$\ddot{x} \equiv (\ddot{z}, \ddot{a})$. For convenience, we also write $r(d, \tilde{x}, \ddot{x}) = r(d, z, a)$.

Proposition 3.1 Suppose Assumptions A.1(a, b) hold such that $E(Y) < \infty$. Then (i) $\tilde{\mu}(D, \tilde{X}) \equiv E(Y \mid D, \tilde{X})$ exists and is finite; and (ii) for each $(d, \tilde{x})$ in $\text{supp}(D, \tilde{X})$ we have

$$\tilde{\mu}(d, \tilde{x}) = \int r(d, \tilde{x}, \ddot{x})\, d\tilde{G}(\ddot{x} \mid d, \tilde{x}). \qquad \blacksquare$$

This simple result represents the conditional expectation $\tilde{\mu}(d, \tilde{x})$ as the *average response given* $(D, \tilde{X}) = (d, \tilde{x})$. For brevity, we call this the "average response," letting the conditioning be implicit. For brevity, we call $\tilde{\mu}$ an *average response function.* This tells us what response to expect given realizations $(d, \tilde{x})$ of $(D, \tilde{X})$ jointly generated by the data generating process specified by A.1(a, b), but it does not say what to expect under interventions to the causes of interest.

3.2.1 Covariate-Conditioned Average Effects of Interventions

When A.1(c) holds, we can define a particular conditional expectation with a clear counterfactual interpretation. Specifically, when $E(r(d, Z, A))$ exists and is finite for each d in $\text{supp}(D)$, we define the *average counterfactual response at d given $\tilde{X} = \tilde{x}$* as

$$\tilde{\rho}(d, \tilde{x}) \equiv E(r(d, Z, A) \mid \tilde{X} = \tilde{x}) = \int r(d, \tilde{x}, \ddot{x})\, d\tilde{G}(\ddot{x} \mid \tilde{x}),$$

where $d\tilde{G}(\cdot \mid \tilde{x})$ denotes the conditional density of $\ddot{X}$ given $\tilde{X} = \tilde{x}$. Assumption A.1(c) ($\mathcal{D} \not\Rightarrow_S \mathcal{Z}$) plays a key role by ensuring different values for d do not necessitate different realizations of Z.

Making this last point more explicit, we view $\tilde{\rho}(d, \tilde{x})$ as representing $\tilde{\rho}(d, \tilde{X}(\omega))$ for $\tilde{X}(\omega) = \tilde{x}$, where $\tilde{X}$ explicitly does not depend on d. In the absence of A.1(c), it is necessary to consider

$$\tilde{\rho}(d, \tilde{X}(d, \cdot)) = E(r(d, Z(d, \cdot), A) \mid \tilde{X}(d, \cdot))$$

making the dependence of Z (hence $\tilde{X}$) on d explicit. This would permit analysis of mediated effects, but we do not anlayze this more complicated case here. For simplicity and notational convenience, whenever A.1(c) holds, we thus interpret $\tilde{\rho}(d, \tilde{x})$ as a representation in which d and $\tilde{x}$ are variation-free without further explicit indication, and similarly for the analogs B.1(c) and ρ of Section 4.

Pearl (see, e.g., Pearl, 1995, 2000) introduced the "do" operator to express counterfactual settings of causes of interest. If the setting of $\mathcal{D}$ is to be d, the expected response given $\tilde{X} = \tilde{x}$ in Pearl's notation is written $E(Y \mid \text{do}(d), \tilde{X} = \tilde{x})$, where the notation do$(d)$ expresses the setting of $\mathcal{D}$ to d. The settable system framework permits this to be straightforwardly represented as $E(Y \mid \text{do}(d), \tilde{X} = \tilde{x}) = \tilde{\rho}(d, \tilde{x})$.

The function $\tilde{\rho}$ is a covariate-conditioned analog of the "average structural function" of Blundell and Powell (2003). Because $\tilde{\rho}(d, \tilde{x})$ tells us what response to expect given $\tilde{X} = \tilde{x}$, for any realized value of d, factual or not, we call $\tilde{\rho}$ a *covariate-conditioned counterfactual average response function.* Leaving conditioning implicit, we will also call $\tilde{\rho}$ a "counterfactual average response function."

Comparing $\tilde{\mu}(d, \tilde{x})$ and $\tilde{\rho}(d, \tilde{x})$, we see that their sole difference is that the density $d\tilde{G}(\ddot{x} \mid d, \tilde{x})$ appears in $\tilde{\mu}(d, \tilde{x})$, whereas $d\tilde{G}(\ddot{x}|\tilde{x})$ appears in $\tilde{\rho}(d, \tilde{x})$. It follows that whenever $d\tilde{G}(\cdot \mid d, \tilde{x}) = d\tilde{G}(\cdot \mid \tilde{x})$, we have $\tilde{\mu}(d, \tilde{x}) = \tilde{\rho}(d, \tilde{x})$, so that the average response does provide counterfactual information. A condition sufficient for this is:

Assumption A.2 The random vector $\ddot{X}$ is independent of D given $\tilde{X}$, written $\ddot{X} \perp D \mid \tilde{X}$. ■

By analogy with common usage of the term "exogeneity" to describe regressors independent of unobservable "disturbances" (e.g., Wooldridge, 2002, p. 50), when A.2 holds we say that the causes of interest D are *conditionally exogenous* given covariates $\tilde{X}$. For conciseness, we refer to this concept as *conditional exogeneity.* As can be readily verified, this concept is distinct from such classical exogeneity concepts such as weak, strong, or super exogeneity (Engle, Hendry, and Richard, 1983). It contains strict exogeneity ($\ddot{X} \perp D$) as a special case (condition on the trivial information set generated by the constant covariate, $\tilde{X} \equiv 1$).

In stating this assumption, we use the conditional independence notation of Dawid (1979). Expressed in terms of conditional densities, Assumption A.2 ensures that for all relevant $(d, \tilde{x}, \ddot{x})$,

$$d\tilde{G}(\ddot{x} \mid d, \tilde{x}) = d\tilde{G}(\ddot{x} \mid \tilde{x}),$$

that is, the conditional density of $\ddot{X}$ given $(D, \tilde{X}) = (d, \tilde{x})$ does not depend on d.

When D is a binary scalar, Assumption A.1 and conditional exogeneity imply Rosenbaum and Rubin's (1983) "unconfoundedness" condition (see White (2005a), proposition 3.2),

$$(Y(0), Y(1)) \perp D \mid \tilde{X},$$

where $Y(0)$ is the response in the absence of treatment ($D = 0$), and $Y(1)$ is the response in the presence of treatment (D =1). (Put $Y(0) = r(\,0,\, Z,\, A)$, $Y(1) = r(\,1,\, Z,\, A)$.) As Rosenbaum and Rubin

(1983) discuss, unconfoundedness ensures the availability of the *propensity score*, $p(\tilde{X}) \equiv P[D = 1 \mid \tilde{X}]$, which has the important property that $(Y(0), Y(1)) \perp D \mid p(\tilde{X})$. For continuous causes D, A.1 and A.2 similarly imply Hirano and Imbens's (2004) "weak unconfoundedness" condition, implying the availability of their *generalized propensity score*, the conditional density of D given $\tilde{X}$.

Conditional exogeneity can be ensured by a generalized form of *randomization*. Randomization, in which $\mathcal{D}$ generates settings D_i randomly and independently of all observed and unobserved attributes and potential causes, was introduced by Fisher (1935, ch.2) as a powerful means of identifying causal effects of interest. For now, we proceed assuming conditional exogeneity. Below we show how this can be ensured by *conditional randomization*.

Our first structural identification result identifies $\tilde{\rho}$ with $\tilde{\mu}$.

Theorem 3.2 Suppose Assumptions A.1(a, b, c) and A.2 hold such that $E(Y) < \infty$. (i) Then for all $(d, \tilde{x}) \in$ supp $(D, \tilde{X})$,

$$\tilde{\rho}(d, \tilde{x}) \equiv \int r(d, \tilde{x}, \ddot{x})\, d\tilde{G}(\ddot{x} \mid \tilde{x})$$

exists, and $\tilde{\rho}(d, \tilde{x}) = \tilde{\mu}(d, \tilde{x})$. ■

This identifies an aspect of the causal structure, the counterfactual average response function $\tilde{\rho}$, with an aspect of the stochastic structure, the conditional expectation $\tilde{\mu}$. Because $\tilde{\mu}$ is empirically meaningful and empirically accessible, so is $\tilde{\rho}$, under the conditions given.

The counterfactual information provided by $\tilde{\rho}$ makes it possible to define the expected effect of any intervention to $\mathcal{D}$. Specifically, consider changing the setting of $\mathcal{D}$ from d to d^*, say. Then the *average effect on $\mathcal{Y}$ of the intervention $d \to d^*$ to $\mathcal{D}$ given $\tilde{X} = \tilde{x}$* is

$$\Delta\tilde{\rho}(d, d^*, \tilde{x}) \equiv \tilde{\rho}(d^*, \tilde{x}) - \tilde{\rho}(d, \tilde{x}).$$

We refer to this as a "covariate-conditioned average effect of intervention," or, leaving the conditioning implicit, an "average effect of intervention." Formally, the interventions (ω, ω^*) underlying $d \to d^*$ given $\tilde{X} = \tilde{x}$ are those satisfying $d = D(\omega)$, $\tilde{x} = \tilde{X}(\omega)$ and $d^* = D(\omega^*)$, $\tilde{x} = \tilde{X}(\omega^*)$.

The immediate consequence of Theorem 3.2(i) is that when A.2 holds, and provided $(d^*, \tilde{x})$ and $(d, \tilde{x})$ both belong to supp $(D, \tilde{X})$, we have

$$\Delta\tilde{\rho}(d, d^*, \tilde{x}) = \Delta\tilde{\mu}(d, d^*, \tilde{x}) \equiv \tilde{\mu}(d^*, \tilde{x}) - \tilde{\mu}(d, \tilde{x}).$$

This provides the fundamental and necessary link between the counterfactual object $\Delta\tilde{\rho}$, which carries the information about causal effects, and the empirically meaningful $\Delta\tilde{\mu}$. Whenever $\Delta\tilde{\mu}(d, d^*, \tilde{x})$ can be consistently estimated, then so can any effect $\Delta\tilde{\rho}(d, d^*, \tilde{x})$.

Replacing $\tilde{x}$ with $\tilde{X}$ yields a random version of the average effect of intervention, $\Delta\tilde{\rho}(d, d^*, \tilde{X})$, that has an optimal prediction property. Specifically, the mean-square optimality property of conditional expectation implies that $\Delta\tilde{\rho}(d, d^*, \tilde{X})$ is the mean-squared-error best predictor of the random effect $\Delta r(d, d^*, Z, A) \equiv r(d^*, Z, A) - r(d, Z, A)$ of the intervention $d \to d^*$, for any such intervention:

$$E([\Delta r(d, d^*, Z, A) - \Delta\tilde{\rho}(d, d^*, \tilde{X})]^2) \leq E([\Delta r(d, d^*, Z, A) - f(d, d^*, \tilde{X})]^2)$$

for all functions $(d, d^*) \to f(d, d^*, \tilde{X})$, since $\Delta\tilde{\rho}(d, d^*, \tilde{X}) = E(\Delta r(d, d^*, Z, A) \mid \tilde{X})$.

Structural identification has important, fundamental implications for the interpretation of regression coefficients. For concreteness and simplicity, consider the familiar case of linear regression, $E(Y \mid D, \tilde{X}) = D\beta^* + \tilde{X}'\gamma^*$, with D a scalar binary indicator (dummy) variable. In the absence of structural identification, β^* and γ^* have only a predictive interpretation, as they deliver mean squared error-optimal predictions of Y given D and X (e.g., White, 1980). Given structural identification, however, β^* acquires a causal interpretation. Specifically, let $d = 0$ and $d^* = 1$. Then for all $\tilde{x}$, $\Delta\mu(d, d^*, \tilde{x}) = \beta^*$; structural identification further ensures $\Delta\rho(d, d^*, \tilde{x}) = \Delta\mu(d, d^*, \tilde{x}) = \beta^*$.

This causal interpretation substantially refines the usual textbook ceteris paribus interpretation for regression coefficients. Specifically, β^* is the covariate-conditioned average effect on the response of an intervention $d = 0 \to d^* = 1$. That is, it is the expected effect of the intervention on the response, *averaging over* the unobserved explanatory variables, *conditional on* the observed covariates. The unobserved explanatory variables are not "held constant," but are averaged over; and the observed covariates are not "held constant," but are conditioned on. These distinctions are especially important, as "holding constant" is a counterfactual operation associated with interventions, whereas averaging and conditioning are stochastic operations. Further, the remaining coefficients γ^* do not have any necessary causal interpretation; they possess only their predictive interpretation.

To gain further insight, let D comprise two dummy variables, $D = (D_1, D_2)$, and suppose $E(Y \mid D, \tilde{X}) = D_1\beta_1^* + D_2\beta_2^* + \tilde{X}'\gamma^*$. Now let $d = (0, 0)$ and $d^* = (1,0)$, so that the intervention changes only d_1, holding d_2 constant at 0. Under structural identification, for all $\tilde{x}$, $\Delta\rho(d, d^*, \tilde{x}) = \Delta\mu(d, d^*, \tilde{x})$ $= \beta_1^*$. We interpret this as the expected effect of an intervention to $d_1(0 \to 1)$ holding d_2 constant (at 0), averaging over the unobserved explanatory variables, conditional on the observed covariates. Now one of the causes of interest (D_2) is held constant. As before, the unobserved explanatory variables are averaged over, not held constant; the observed covariates are conditioned on, not held constant. Here β_1^* and β_2^* have causal interpretations, but γ^* only has a predictive interpretation.

These interpretive considerations hold generally. In the context of linear regression, the upshot is that some coefficients have causal interpretations, and others do not. Recognition of this fact could have far-reaching implications for the way researchers assess the plausibility of their empirical results.

Similarly, in more general contexts, certain differences or derivatives have causal meaning, while others do not. Which these are is determined by structural identification.

3.2.2 Covariate-Conditioned Average Marginal Effects

Another effect that is often of interest is the *average marginal ceteris paribus effect on $\mathcal{Y}$ of $\mathcal{D}_j$ given* $(D, \tilde{X}) = (d, \tilde{x})$, which we write

$$\tilde{\xi}_j(d, \tilde{x}) \equiv \int (\partial r / \partial d_j)(d, \tilde{x}, \ddot{x}) \, d\tilde{G}(\ddot{x} \mid d, \tilde{x}),$$

provided r is suitably differentiable and the indicated integral exists. This is related to the average derivatives considered by Stoker (1986) and Powell, Stock, and Stoker (1989).

This is a particular weighted average of the unobservable marginal effects $(\partial r/\partial d_j)(d, z, a)$, averaging over both unobserved causes and essential heterogeneity, given observed causes and attributes. Note that we include d in the conditional density appearing in the defining expression. This later permits us to straightforwardly examine both necessary and sufficient conditions for the identification of this effect. We call $\tilde{\xi}_j(d, \tilde{x})$ a "covariate-conditioned average marginal effect," or, leaving conditioning implict, an "average marginal effect."

The next regularity condition, permitting interchange of derivative and integral in the representation of $\tilde{\rho}$ in Theorem 3.2(i), delivers structural identification of average marginal effects. To state this, let supp $(\ddot{X} \mid \tilde{x})$ denote the conditional support of $\ddot{X}$ given $\tilde{X} = \tilde{x}$. That is, supp $(\ddot{X} \mid \tilde{x})$ is the smallest measurable set on which $d\tilde{G}(\cdot \mid \tilde{x})$ integrates to 1. We let d_j denote the jth element of d and let $d_{(j)}$ denote the $(k_1 - 1)$-vector containing all but d_j, $j \in \{1, \ldots, k_1\}$.

Assumption A.3 For given $(d, \tilde{x}) \in$ supp $(D, \tilde{X})$, suppose the function $\ddot{x} \to r(d, \tilde{x}, \ddot{x})$ is integrable with respect to $\tilde{G}(\cdot \mid \tilde{x})$, that is,

$$\int r(d, \tilde{x}, \ddot{x}) \, d\tilde{G}(\ddot{x} \mid \tilde{x}) < \infty,$$

and suppose that $(\partial r/\partial d_j)(d, \tilde{x}, \ddot{x})$ exists on $C_j\times$ supp $(\ddot{X} \mid \tilde{x})$, where C_j is a convex compact neighborhood of the given d_j, for the given $(d_{(j)}, \tilde{x})$. Suppose further that for the given $(d_{(j)}, \tilde{x})$ and for each $\ddot{x}$ in supp $(\ddot{X} \mid \tilde{x})$,

$$\sup_{d_j \in C_j} \mid (\partial r/\partial d_j)(d, \tilde{x}, \ddot{x}) \mid \leq q(d_{(j)}, \tilde{x}, \ddot{x}),$$

where q is a measurable function such that

$$\int q(d_{(j)}, \tilde{x}, \ddot{x})\, d\tilde{G}(\ddot{x} \mid \tilde{x}) < \infty \qquad \blacksquare$$

As a convenient shorthand, when the conditions of A.3 hold, we say that "$(\partial r/\partial d_j)(d, \tilde{x}, \ddot{x})$ is dominated on C_j by a function integrable with respect to $\tilde{G}(\cdot \mid \tilde{x})$ at $(d, \tilde{x})$."

We state our next structural identification result as a continuation of Theorem 3.2:

Theorem 3.2(ii) If, in addition to the conditions of Theorem 3.2(i), Assumption A.3 holds, then the functions $d_j \to \tilde{\rho}(d, \tilde{x})$ and $d_j \to \tilde{\mu}(d, \tilde{x})$ are differentiable on C_j, and

$$\begin{aligned} (\partial\tilde{\mu}/\partial d_j)(d, \tilde{x}) &= (\partial\tilde{\rho}/\partial d_j)(d, \tilde{x}) \\ &= \tilde{\xi}_j(d, \tilde{x}) \equiv \int (\partial r/\partial d_j)(d, \tilde{x}, \ddot{x})\, d\tilde{G}(\ddot{x} \mid d, \tilde{x}). \qquad \blacksquare \end{aligned}$$

This identifies an aspect of the causal structure, the average marginal effect of interest $\tilde{\xi}_j$, with an aspect of the stochastic structure, the partial derivative with respect to d_j of $\tilde{\mu}$. Consequently, to consistently estimate $\tilde{\xi}_j$, it suffices to consistently estimate this derivative. Note that identification here is for given j, point-wise in $(d, \tilde{x})$. That is, identification is guaranteed for any j and $(d, \tilde{x})$ for which Assumption A.3 holds, but is not necessarily guaranteed otherwise.

General conditions under which relevant derivatives can be consistently estimated are well known. Chalak and White (2006) give conditions suitable for the present context.

The average marginal effect is identified only for those causes for which conditional exogeneity is ensured, that is, D. The derivatives with respect to the observed ancillary causes $\tilde{Z}$ (for which conditional exogeneity need not hold) do *not* have any necessary causal interpretation. Further, the derivatives with respect to the observed attributes $\tilde{A}$ *cannot* have any causal interpretation, as attributes are not subject to intervention and therefore cannot be causes.

We now show that conditional exogeneity holds under a generalization of randomization.

Proposition 3.3 Given Assumption A.1(a), suppose $D = c(\tilde{X}, U)$, where c is a measurable function and U is a random vector such that $\ddot{X} \perp U \mid \tilde{X}$. Then $\ddot{X} \perp D \mid \tilde{X}$, that is, A.2 holds. ■

Pure randomization occurs when $D = U$ and U is independent of $\tilde{X}$ and $\ddot{X}$. (This is sufficient for $\ddot{X} \perp U \mid \tilde{X}$ by Dawid (1979, lemma 4.2(ii)).) We also have that randomization dependent on $\tilde{X}$, i.e., *conditional randomization*, ensures conditional exogeneity. The next corollary is immediate:

Corollary 3.4 Suppose Assumptions A.1(a, b, c) hold and that $D = c(\tilde{X}, U)$, where c is a measurable function and U is a random vector such that $\ddot{X} \perp U \mid \tilde{X}$. Then the conclusions of Theorem 3.2(i) hold. If in addition A.3 holds, then the conclusions of Theorem 3.2(ii) hold. ■

Thus, when the researcher controls the generation of D, then structural identification of average effects of interest can be ensured using conditional randomization.

Note that nothing here rules out always setting two elements of D to the identical values. This will make it impossible in applications to separate out the two effects of interest, but the problem with this is a lack of *stochastic* identification, rather than a lack of structural identification.

The control necessary to implement conditional randomization is common for experimental or clinical researchers, and use of experimental methods to study economic behavior is a rapidly expanding field (see, e.g. Lucking-Reiley, 1999; List and Lucking-Reiley, 2000, 2002; Fershtman and Greezy, 2001; Harrison, Lau and Williams, 2002; Karlan, 2003; Eckel, Johnson, and Montmarquette, 2003; List, 2004; and Bettinger and Slonin, 2004). The results of this section are thus directly relevant there. Outside the laboratory, economists can sometimes exploit randomization (e.g., Angrist, 1998), but in the absence of experimental control, economists typically cannot rely on A.2 to identify effects of interest. We study the consequences of the failure of A.2 as a special case of the more general approach to identifying effects of interest given in Section 4.

3.2.3 More General Measures of Effect

Although substantial interest may attach to the average effects discussed above, interest may also attach to effects of interventions on other aspects of the conditional distribution of the response. Heckman, Smith, and Clements (1997) draw attention to this issue in the context of programme evaluation. Imbens and Newey (2003) discuss a variety of such effects. In the context of wage determination, Firpo, Fortin, and Lemieux (2005) develop methods to study average effects of binary treatments on general aspects of the unconditional response distribution, such as the variance, median, or density. In this section we discuss how these methods can be extended to identify causal effects of general interventions on general aspects of the conditional distribution of the response.

A simple way to define general effects uses the *covariate-conditioned counterfactual moment*

$$\tilde{\rho}_0(d, \tilde{x}) \equiv \tau_0(\tilde{\rho}_1(d, \tilde{x}), \tilde{\rho}_2(d, \tilde{x}), \ldots),$$

where τ_0 is a known function, and for known scalar-valued functions τ_k,

$$\tilde{\rho}_k(d, \tilde{x}) \equiv \int \tau_k(r(d, \tilde{x}, \ddot{x}))\, d\tilde{G}(\ddot{x} \mid \tilde{x}), \qquad k = 1, 2, \ldots,$$

provided integrals exist. The *moment effect on $\mathcal{Y}$ of the intervention $d \to d^*$ to $\mathcal{D}$ given $\tilde{X} = \tilde{x}$* is

$$\Delta\tilde{\rho}_0(d, d^*, \tilde{x}) \equiv \tilde{\rho}_0(d^*, \tilde{x}) - \tilde{\rho}_0(d, \tilde{x}),$$

and the *marginal moment effect on $\mathcal{Y}$ of $\mathcal{D}_j$ given $\tilde{X} = \tilde{x}$* is $(\partial\tilde{\rho}_0/\partial d_j)(d, \tilde{x})$.

As a simple example, let $\tau_1(r) = 1[r \leq y]$ for $y \in \mathbb{R}$ (cf. Imbens, 2004, p.9), and let $\tau_0(\tilde{\rho}_1) = \tilde{\rho}_1$. Then effects on the conditional distribution of the response are defined from

$$\tilde{\rho}_0(d, \tilde{x}) = \int 1[r(d, \tilde{x}, \ddot{x}) \leq y]\, d\tilde{G}(\ddot{x} \mid \tilde{x}),$$

which defines the counterfactual conditional distribution function of the response. As a somewhat more elaborate example, let $\tau_1(r) = r$, $\tau_2(r) = r^2$, and put

$$\tilde{\rho}_0(d, \tilde{x}) = \tilde{\rho}_2(d, \tilde{x}) - \tilde{\rho}_1(d, \tilde{x})^2.$$

This defines the covariate-conditioned counterfactual variance function, from which effects on the conditional variance of the response follow.

When conditional exogeneity holds, the counterfactual moment function and the corresponding effects are identified, because then $\tilde{\rho}_0(d, \tilde{x}) = \tilde{\mu}_0(d, \tilde{x})$, where

$$\begin{aligned} \tilde{\mu}_0(d, \tilde{x}) &\equiv \tau_0(\tilde{\mu}_1(d, \tilde{x}), \tilde{\mu}_2(d, \tilde{x}), \ldots) \\ \tilde{\mu}_k(d, \tilde{x}) &\equiv \int \tau_k(r(d, \tilde{x}, \ddot{x}))\, d\tilde{G}(\ddot{x} \mid d, \tilde{x}), \qquad k = 1, 2, \ldots \qquad . \end{aligned}$$

Another way to define general effects uses the covariate-conditioned counterfactual optimizer

$$\tilde{\rho}_0(d, \tilde{x}) \equiv \arg\max_m \int \tau(r(d, \tilde{x}, \ddot{x}), m)\, d\tilde{G}(\ddot{x} \mid \tilde{x}),$$

where $\tau : \mathbb{R} \times \mathbb{R}^{\lambda} \to \mathbb{R}$ is a known function, so that $\tilde{\rho}_0(d, \tilde{x})$ is a $\lambda \times 1$ vector of aspects of the counterfactual conditional distribution of the response determined by choice of τ. For example, effects on the conditional α-quantiles of the response can be defined by taking

$$\tau(r, m) = -\left|r - m\right| (\alpha 1[r \geq m] + (1 - \alpha) 1[r < m]).$$

In this case $\tilde{\rho}_0(d, \tilde{x})$ defines the covariate-conditioned counterfactual α-quantile function, a covariate-conditioned analog of the "quantile structural function" of Imbens and Newey (2003). The associated effect is the covariate-conditioned analog of the quantile treatment effect of Lehmann (1974) and Abadie, Angrist, and Imbens (2002).

By taking m to be a vector and taking $\tau(r, m)$ to define a quasi-log-likelihood function, the optimization approach can focus attention simultaneously on multiple aspects of the counterfactual conditional response distribution, such as location and scale generally or other aspects of conditional distribution shape. Taking $\tau(r, m)$ to define a utility function instead of a quasi-log-likelihood function defines effects on optimal actions of interventions to d conditional on $\tilde{x}$. Viewed this way, an optimal action is just an aspect of the conditional distribution of the response.

Yet another way to define effects is based on implicitly defined moments $\tilde{\rho}_0(d, \tilde{x})$ such that

$$\int \tau(r(d, \tilde{x}, \ddot{x}), \tilde{\rho}_0(d, \tilde{x}))\, d\tilde{G}(\ddot{x} \mid \tilde{x}) = 0.$$

Here $\tau : \mathbb{R} \times \mathbb{R}^{\lambda} \to \mathbb{R}^{\lambda}$ is a known function. This contains many instances of the moment or optimizer approaches above as special cases. For example, this implicit definition for $\tilde{\rho}_0(d, \tilde{x})$ can represent first order conditions defining the interior optimizer of some objective function. The implicit moment approach generalizes and complements the optimizer approach in the same way that method of moments estimation generalizes and complements maximum likelihood estimation.

For brevity, we defer formal statement of identification results for these covariate-conditioned counterfactual moment and distributional aspect functions until Section 4.

4 Structural Identification with Predictive Proxies

4.1 Structural Identification

In the treatment effects literature, it is well known that even without randomization the availability of suitably behaved covariates can be exploited to identify average effects of interest. The availability of suitable covariates is typically simply assumed, but the literature provides little guidance about their construction. The following assumption introduces additional covariates that permit straightforward analysis of conditions ensuring structural identification. This also yields significant insight into the nature of these additional covariates.

Assumption B.1 Let an attribute-indexed recursive settable system $\mathcal{S}$ generate a sample from $\mathbf{A} \subset \mathcal{A}$ involving settable variables ($\mathcal{Y}$, $\mathcal{D}$, $\mathcal{W}$, $\mathcal{Z}$) such that $\mathcal{Y} \not\Rightarrow_{\mathcal{S}} (\mathcal{D}, \mathcal{W}, \mathcal{Z})$. In addition:

(a) Let attributes $\{(A_i, \tilde{B}_i) \equiv (\tilde{A}_i, \ddot{A}_i, \tilde{B}_i)\}$ be a sequence of random vectors, and let $(\mathcal{D}, \mathcal{W}, \mathcal{Z})$ generate settings $\{(D_i, W_i, Z_i) \equiv (D_i, W_i, \tilde{Z}_i, \ddot{Z}_i)\}$ such that the joint distribution of $(D_i, X_i) \equiv (D_i, W_i, \tilde{Z}_i, \tilde{A}_i, \tilde{B}_i)$ is H and the conditional distribution of $\ddot{X}_i \equiv (\ddot{Z}_i, \ddot{A}_i)$ given $(D_i, X_i) = (d, x)$ is $G(\cdot \mid d, x)$ for all $i = 1, 2, \ldots$, where D_i is $\mathbb{R}^{k_1}$-valued, k_1 a positive integer, W_i is $\mathbb{R}^{k_2}$-valued, $k_2 \in \mathbb{N}$, $\tilde{Z}_i$ is $\mathbb{R}^{k_3}$-valued, $k_3 \in \mathbb{N}$, $\ddot{Z}_i$ is $\mathbb{R}^{\infty}$-valued, $\tilde{A}_i$ is ℓ_1-valued, $\ell_1 \in \mathbb{N}$, $\ddot{A}_i$ is $\mathbb{R}^{\infty}$-valued, and $\tilde{B}_i$ is ℓ_2-valued, $\ell_2 \in \mathbb{N}$;

(b) The responses $\{Y_i\}$ of $\mathcal{Y}$ are determined as

$$Y_i = r(D_i, Z_i, A_i), \qquad i = 1, 2, \ldots,$$

where r is an unknown measurable scalar-valued function;

(c) (i) $\mathcal{D} \not\Rightarrow_{\mathcal{S}} (\mathcal{W}, \mathcal{Z})$; (ii) $\mathcal{W} \not\Rightarrow_{\mathcal{S}} \mathcal{Z}$

(d) The realizations of $Y_i, D_i, W_i, \tilde{Z}_i, \tilde{A}_i$ and $\tilde{B}_i$ are observed; those of $\ddot{Z}_i$ and $\ddot{A}_i$ are not. ■

This resembles A.1, but it introduces two new objects: observable settable variables $\mathcal{W}$ and observable attributes $\tilde{B}$, both finitely dimensioned. These variables are structurally irrelevant to the response, as, according to B.1(b), $\mathcal{W}$ does not cause $\mathcal{Y}$ and $\tilde{B}$ does not modify the response.

Despite their structural irrelevance, W and $\tilde{B}$ have predictive relevance. Specifically, when W and $\tilde{B}$ proxy for the unobservable $\ddot{X}$, they can help predict the response better than otherwise possible. As we will see, the availability of W and $\tilde{B}$ can help ensure structural identification.

Given their role as proxies for unobservables and their predictive relevance, we refer to $\mathcal{W}$ and $\tilde{B}$ as *predictive proxies*. We call W "proxy settings" and $\tilde{B}$ "proxy attributes." The covariates $X \equiv (W, \tilde{Z}, \tilde{A}, \tilde{B})$ contain observable ancillary causes, relevant attributes, and predictive proxies.

The fact that $\mathcal{W}$ generates settings W used to predict the response necessitates restrictions analogous to those imposed on $\mathcal{Z}$: $\mathcal{Y}$ does not cause $\mathcal{W}$, and $\mathcal{D}$ does not cause $\mathcal{W}$ (B.1(c.i)). Otherwise $\mathcal{W}$ may mediate effects of $\mathcal{D}$. Because $\tilde{B}$ is an attribute, it cannot mediate effects. In B.1(c.ii), we impose some further causal structure, examined in detail below.

The conditional exogeneity condition is now

Assumption B.2 $\ddot{X} \perp D \mid X$. ■

Similar to the situation above, B.1 and B.2 (conditional exogeneity) imply Rosenbaum and Rubin's (1983) unconfoundedness or Hirano and Imbens's (2004) weak unconfoundedness, depending on whether D is binary or continuous. Thus, when B.1 and B.2 hold, X is a set of *sufficient* covariates in Dawid's (1979) terminology.

We now state structural identification results for the general measures of effect introduced in Section 3.2.3, letting $dG(\ddot{x} \mid x)$ denote the conditional density of $\ddot{X}$ given $X = x$, $x \equiv (w, \tilde{z}, \tilde{a}, \tilde{b})$.

Theorem 4.1 Suppose Assumptions B.1(a, b) hold. For $k = 1, 2, \ldots$, let $\tau_k : \mathbb{R} \to \mathbb{R}$ be a measurable function such that $E(\tau_k(Y)) < \infty$. (i) Then $\mu_k(D, X) \equiv E(\tau_k(Y) \mid D, X)$ exists and is finite, and for each $(d,\ x)$ in supp$(D,\ X\ \ )$

$$\mu_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x}))\ dG(\ddot{x} \mid d, x), \qquad k = 1, 2, \ldots.$$

If $\tau_0 : \mathbb{R}^\infty \to \mathbb{R}$ is a measurable function, then the function μ_0 defined by

$$\mu_0(d, x) = \tau_0(\mu_1(d, x), \mu_2(d, x), \ldots)$$

is also measurable. (ii) If B.1(c.i) and B.2 also hold, then for each $(d,\ x)$ in supp$(D,\ X\ \ )$

$$\rho_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x}))\ dG(\ddot{x} \mid x)$$

exists and is finite, and $\rho_k = \mu_k$, $k = 1, 2, \ldots$; the function ρ_0 defined by

$$\rho_0(d, x) = \tau_0(\rho_1(d, x), \rho_2(d, x), \ldots)$$

is measurable; and $\rho_0 = \mu_0$. ■

Theorem 4.2 Suppose Assumptions B.1(a, b) hold. (i) For $\lambda \in \mathbb{N}$, let $\tau : \mathbb{R} \times \mathbb{R}^\lambda \to \mathbb{R}$ be a measurable function such that $E(\tau(Y, m)) < \infty$ for each m in $\mathbb{R}^\lambda$. Then for each (d, x, m) in $\text{supp}(D, X) \times \mathbb{R}^\lambda$ the conditional expectation

$$\varphi_{\tau,d}(d, x, m) = \int \tau(r(d, \tilde{x}, \ddot{x}), m)\ dG(\ddot{x} \mid d, x)$$

exists and is finite. (ii) Further, let τ, r, and (d, x) $\to G(\cdot \mid d, x)$ be such that $\varphi_{\tau,d}(d, x, m)$ defines a continuous real-valued function on $\text{supp}(D, X) \times \mathbb{R}^\lambda$, and let M: $\text{supp}(D, X) \to \mathbb{R}^\lambda$ be a non-empty and compact-valued continuous correspondence. Then for each (d, x) in supp (D, X) the correspondence

$$\mu_0(d, x) = \arg \max_{m \in M(d,x)} \varphi_{\tau,d}(d, x, m)$$

is non-empty, compact-valued, and upper hemi-continuous. (iii) If B.1(c.i) and B.2 hold also, then

$$\varphi_\tau(d, x, m) = \int \tau(r(d, \tilde{x}, \ddot{x}), m)\ dG(\ddot{x} \mid x)$$

defines a continuous real-valued function on supp(D, X) $\times\ \mathbb{R}^\lambda$ such that for each (d, x, m) in supp(D, X)$\times\ \mathbb{R}^\lambda$ we have $\varphi_\tau(d, x, m) = \varphi_{\tau,d}(d, x, m)$; the correspondence ρ_0 defined by

$$\rho_0(d, x) = \arg \max_{m \in M(d,x)} \varphi_\tau(d, x, m)$$

is non-empty, compact-valued, and upper hemi-continuous; and $\rho_0 = \mu_0$. ■

Theorem 4.3 Suppose Assumptions B.1(a, b) hold. (i) For $\lambda \in \mathbb{N}$, let $\tau : \mathbb{R} \times \mathbb{R}^\lambda \to \mathbb{R}^\lambda$ be a measurable function such that $E(\tau(Y, m)) < \infty$ for each $m \in M \subset \mathbb{R}^\lambda$. Then for each ($d$, x, m) in supp(D, X) $\times$ M the conditional expectation

$$\varphi_{\tau,d}(d, x, m) = \int \tau(r(d, \tilde{x}, \ddot{x}), m)\ dG(\ddot{x} \mid d, x)$$

exists and is finite. (ii) Further, let τ, r, and $(d, x) \to G(\cdot \mid d, x)$ be such that for each $(d,\ x,\ m)$ in $\text{supp}(D,\ X) \times M$, $\varphi_{\tau,d}$ is differentiable on a neighborhood of $(d,\ x,\ m)$, the $\lambda \times \lambda$ matrix $\nabla_m \varphi_{\tau,d}(d,\ x,\ m)$ is non-singular, and $\varphi_{\tau,d}(d, x, m) = 0$. Then there exists a unique function μ_0 such that for each $(d,\ x) \in \text{supp}(D,\ X)$, μ_0 is differentiable at $(d,\ x)$ and

$$\int \tau(r(d, \tilde{x}, \ddot{x}), \mu_0(d, x))\ dG(\ddot{x} \mid d, x) = 0.$$

(iii) If B.1(c.i) and B.2 hold also, then there exists a unique function ρ_0 such that for each $(d,\ x) \in \text{supp}(D,\ X)$, ρ_0 is differentiable at $(d,\ x)$;

$$\int \tau(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x))\ dG(\ddot{x} \mid x) = 0;$$

and $\rho_0 = \mu_0$. ■

These results identify the covariate-conditioned counterfactual moment or distributional aspect functions ρ_0 with the empirically accessible covariate-conditioned moment or aspect functions μ_0. (When we write $\rho_0 = \mu_0$, we mean that $\rho_0(d, x) = \mu(d, x)$ for all (d, x) in a subset of supp (D, X) having probability one.) In each case, the non-causality condition $D \not\Rightarrow|_S (\mathcal{W}, \mathcal{Z})$ (B.1(c.i)) and conditional exogeneity (B.2) play the key roles in delivering structural identification. The non-causality condition ensures that any responses represented by elements of x do not depend on d, so changes to d do not result in changes to x. Conditional exogeneity ensures that the covariates X supply sufficient information for predicting $\ddot{X}$ that D has no further predictive value.

For $(d,\ x)$ and (d^*, x) in $\text{supp}(D,\ X\ )$ and for given τ, the *moment* or *distributional aspect effect of the intervention $d^* \to d$ to $\mathcal{D}$ given $X = x$ is*

$$\Delta\rho_0(d, d^*, x) \equiv \rho_0(d^*, x) - \rho_0(d, x).$$

The interventions (ω, ω^*) underlying $d \to d^*$ given $X = x$ are those satisfying $d = D(\omega)$, $x = X(\omega)$ and $d^* = D(\omega^*)$, $x = X(\omega^*)$. The results above identify these effects as

$$\Delta\rho_0(d, d^*, x) = \Delta\mu_0(d, d^*, x) \equiv \mu_0(d^*, x) - \mu_0(d, x).$$

Observe that the value x need not be factual, so not only is there a necessary counterfactual aspect to the intervention, there is also an allowed counterfactual aspect to the conditioning. Comparing the effects of an intervention $d^* \to d$ given different values of X, say x and x^*, gives an *average effect difference* $\Delta\rho_0(d, d^*, x^*) - \Delta\rho_0(d, d^*, x)$ measuring how the expected effect would be modified by a "change" in the covariate outcome. The covariates are thus often referred to in the epidemiological literature as "effect modifiers," an accurate and useful description.

Like $\Delta\tilde{\rho}_0$, $\Delta\rho_0$ is a covariate-conditioned effect, but whereas $\Delta\tilde{\rho}_0$ pays attention only to $\tilde{X}$, $\Delta\rho_0$ takes account of X, which includes not only $\tilde{X}$, but also the predictive proxies. The use of suitable predictive proxies can thus yield more accurate predictors of the underlying random effect.

In Section 3.2, we treated identification of both non-marginal and marginal effects. We defer treating marginal effects to Section 4.4, where we study weaker conditions for structural identification.

4.2 Constructing Covariates

We now investigate the construction of the covariates using the following analog of Proposition 3.3.

Proposition 4.4 Given Assumption B.1(a), suppose $D = c(X,\ U)$, where c is a measurable function and U is a random vector such that $\ddot{X} \perp U \mid X$. Then $\ddot{X} \perp D \mid X$, that is, B.2 holds. ■

The analog of Corollary 3.4 is

Corollary 4.5 Suppose Assumptions B.1(a, b, c.i) hold and that $D = c(X,\ U)$, where c is a measurable function and U is a random vector such that $\ddot{X} \perp U \mid X$. If the other assumptions of Theorems 4.1, 4.2, or 4.3 hold, then the conclusions of those results also hold. ■

Proposition 4.4 says that conditional exogeneity holds provided treatment can be expressed as a response depending only on (observable) covariates X and variables U independent of the unobservable explanatory variables $\ddot{X}$, given the covariates. This response could be under researcher control, extending the notion of conditional randomization to include treatment assignment based on covariates X, rather than just observable structurally relevant explanatory variables $\tilde{X}$.

Alternatively, D may be a response beyond researcher control. For example, agents may self-select treatment, and if this self-selection can be represented in terms of (observable) X and (unobservable) U with the given properties, then conditional exogeneity holds.

To gain insight, consider that when D is a response beyond researcher control, we have

$$D = \ddot{c}(\tilde{X}^*, \ddot{X}^*),$$

for some unknown measurable function $\ddot{c}$ and "D-relevant" explanatory variables $(\tilde{X}^*, \ddot{X}^*)$, say, where $\tilde{X}^*$ is observable and $\ddot{X}^*$ is not. The D-relevant explanatory variables may include elements of (Z, A); they may also include elements not belonging to Z or A. In the latter case, the settable variables generating the non-attribute components of $(\tilde{X}^*, \ddot{X}^*)$ cannot be caused by $\mathcal{Y}$, as this would violate the requirement that $\mathcal{Y} \not\Rightarrow_S \mathcal{D}$; nor can these be caused by $\mathcal{D}$, as this would violate B.1(c.i).

Examining this expression for D, we see that the conditions of Proposition 4.3 hold with $c = \ddot{c}$, $X = \tilde{X}^*$, and $U = \ddot{X}^*$, for the special case in which $\ddot{X} \perp \ddot{X}^* \mid \tilde{X}^*$, that is, the case in which given $\tilde{X}^*$, $\ddot{X}^*$ provides no further predictive information for $\ddot{X}$. In this case, the covariates are simply $X = \tilde{X}^*$.

This is indeed a special case, as we might expect $\ddot{X}^*$ to contain useful predictive information for $\ddot{X}$, even given $\tilde{X}^*$, as $\ddot{X}^*$ and $\ddot{X}$ could contain common elements. We proceed, therefore, by augmenting $\tilde{X}^*$ with additional observable variables $\tilde{X}^+$, say, sufficiently predictive for $\ddot{X}$ so that given $X = (\tilde{X}^*, \tilde{X}^+)$, $\ddot{X}^*$ no longer contains useful predictive information for $\ddot{X}$: that is, we have $\ddot{X} \perp \ddot{X}^* \mid X$. Moreover, the presence of the "D-irrelevant" variables $\tilde{X}^+$ does not have adverse consequences for the representation of the response D, as we formally have

$$D = c(\tilde{X}^*, \tilde{X}^+, \ddot{X}^*) = \ddot{c}(\tilde{X}^*, \ddot{X}^*),$$

where the role of the function c is to "ignore" the D-irrelevant variables. The additional variables $\tilde{X}^+$ thus permit us to satisfy the conditions of Proposition 4.4 with $U = \ddot{X}$. An obvious source of such variables is $\tilde{X}$, as $\ddot{X}$ and $\tilde{X}$ are explicitly permitted to be dependent under B.2.

What other sources of predictors for $\ddot{X}$ might there be? To proceed further, it is helpful to consider separately predictors for attributes and predictors for causes. For attributes $\ddot{A}$, useful proxies are observable attributes other than $\tilde{A}$, say $\tilde{B}$, correlated with $\ddot{A}$. (A more precise phrase than "correlated with" is "statistically dependent on," but we informally use "correlated," as it is less awkward.) For example, the attribute of height could be used as a proxy for the attribute of gender and vice-versa.

In considering proxies for $\ddot{Z}$, it is useful to apply Reichenbach's (1956) *principle of common causality*, as in White (2005a). This principle holds that dependence between two potential causes arises either because one variable is the cause of the other or because the two variables share a common cause. Here, this maps to a situation in which one settable variable responds to the other or both respond to a common third settable variable. As explained in White (2005a), it follows that the only relevant channel for the dependence sought here is for the observable proxy to be a response to $\ddot{Z}$.

Specifically, if $\ddot{Z}$ were instead a response to a potential proxy, then a suitable substitution in the response function r yields a modified response function containing the potential proxy as a structurally relevant observable cause. Although this may be helpful, this produces an observable cause, not a proxy for an unobservable cause. Or if both $\ddot{Z}$ and the observable proxy respond to a third settable variable (unobservable, without loss of generality), then substituting the response function for $\ddot{Z}$ into r yields a proxy that is again a response to an unobservable cause, the underlying common cause. Thus, we seek predictive proxies that are responses to unobservable structurally relevant variables $\ddot{Z}$.

Measurement-error laden versions of $\ddot{Z}$ are useful proxies of this sort. Measurement-error laden versions of unobserved structurally relevant attributes are also useful proxy attributes. (These latter variables are no longer attributes when the measurement error is not itself an attribute.)

A similar line of reasoning begins by noting that $\ddot{X} \perp \ddot{X}^* \mid X$ also implies the symmetrical requirement that additional observable variables $\tilde{X}^+$ are sufficiently predictive for $\ddot{X}^*$ so that given $X = (\tilde{X}^*, \tilde{X}^+)$, $\ddot{X}$ no longer contains useful predictive information for $\ddot{X}^* \equiv (\ddot{Z}^*, \ddot{A}^*)$. Parallel to the discussion above, one may seek observable attributes $\tilde{B}$ correlated with the unobservable D-relevant attributes, say $\ddot{A}^*$. Similarly, one may seek predictive proxies dependent on the unobservable D-relevant causal factors $\ddot{Z}^*$. Parallel to the above, the only relevant channel for the dependence sought here is for the observable proxy to be a response to $\ddot{Z}^*$. Just as above, measurement error-laden versions of $\ddot{Z}^*$ are useful of this sort.

Summarizing, the covariates should include: (1) observable D-relevant explanatory variables $\tilde{X}^*$; (2) observable structurally relevant explanatory variables $\tilde{X}$; (3) observable attributes $\tilde{B}$ correlated with unobservable structurally and D-relevant attributes $\ddot{A}$ and $\ddot{A}^*$; and (4) observable responses to unobservable structurally relevant and D-relevant causes $\ddot{Z}$ and $\ddot{Z}^*$.

In Assumption B.1, we represent the covariates as $X = (W, \tilde{X}, \tilde{B})$. It follows that the proxy settings W can be constructed as observable responses of the form

$$W = w(\tilde{X}^*, \ddot{X}^*, \ddot{X}, \tilde{X}, \tilde{B}, \ddot{B}, V),$$

where w is a measurable vector-valued function. The functional form of some components of w is known, as W should necessarily include elements of $\tilde{X}^*$ not already present in $\tilde{X}$. Other components obey unknown relationships, but it is sufficient that these are determined by the listed arguments but not D or Y. Arguments not previously defined are $\ddot{B}$, a vector of W-relevant unobservable attributes, and V, a vector of unobservable random variables (settings) such that $V \perp D \mid \tilde{X}^*, \ddot{X}^*, \ddot{X}, \tilde{X}, \tilde{B}, \ddot{B}$.

The latter variables V can be thought of as measurement errors. The conditional independence required of V is a natural one for measurement errors and is necessary to ensure B.2 with $X = (W, \tilde{X}, \tilde{B})$ for $(W, \tilde{B})$ as just given.

It is important to emphasize that economic theory and application domain knowledge play the key roles in identifying legitimate and illegitimate elements of each component of the covariates. Economic theory identifies the structurally relevant explanatory variables, the D-relevant explanatory variables, and the responses yielding the proxy settings W. Application domain knowledge identifies which variables are observable or unobservable, and which observable attributes may be correlated with unobservable structurally or D-relevant attributes. Economic theory also identifies variables caused by $\mathcal{Y}$ or $\mathcal{D}$ that are not legitmate for constructing predictive proxies.

This discussion justifies Assumption B.1(c.ii): $\mathcal{W} \not\Rightarrow|_S \mathcal{Z}$. Thus, the referenced causal variables obey the recursive structure $\mathcal{Y} \not\Rightarrow|_S (\mathcal{D}, \mathcal{W}, \mathcal{Z})$, $\mathcal{D} \not\Rightarrow|_S (\mathcal{W}, \mathcal{Z})$, and $\mathcal{W} \not\Rightarrow|_S \mathcal{Z}$. This is not a complete characterization of the recursive structure of the settable system, as not every cause of $\mathcal{W}$ is referenced by $\mathcal{Z}$, and not every cause of $\mathcal{D}$ is referenced by $(\mathcal{W}, \mathcal{Z})$. The stated conditions nevertheless suffice. Assumption B.1(c.ii) is not necessary for our results; we include it in B.1 to emphasize the natural recursive structure that arises when using predictive proxies.

Although this analysis provides specific guidance not elsewhere available for choosing covariates, constructing predictive proxies $(W, \tilde{B})$ as just described cannot be proven to ensure $\ddot{X} \perp \ddot{X}^* \mid X$, so Corollary 4.5 will not necessarily apply with $U = \ddot{X}^*$ to deliver B.2. Nevertheless, one can test conditional exogeneity for given predictive proxies; see White and Chalak (2006).

4.3 Implications for Estimation

Once any given counterfactual aspect of the conditional response distribution is structurally identified, its estimation reduces to estimating the corresponding standard aspect of the conditional response distribution. Estimates of effects of interest follow by taking suitable differences or derivatives of the estimated standard distributional aspect.

A detailed study of the properties of parametric and nonparametric estimators of these objects is taken up in our companion paper (White and Chalak, 2006) in a setting most suitable for cross-section data. White (2005a) considers estimation of effects of certain binary causes in a time series setting. Nevertheless, to make the process fully explicit, we now sketch the construction of estimators for counterfactual objects under structural identification.

First, consider the covariate-conditioned optimizer of Theorem 4.2,

$$\mu_0(d,x) \equiv \arg\max_m \int \tau(r(d,\tilde{x},\ddot{x}),m)\ \mathrm{d}G(\ddot{x} \mid d,x).$$

The covariate-conditioned counterfactual optimizer ρ_0 equals μ_0 given structural identification, so it suffices to estimate μ_0. To do this, "parameterize" the argument m above by specifying a function of parameters θ, say $\theta \to m(d,x,\theta)$, such that $m(d,x,\theta^*) = \mu_0(d,x)$ for all (d,x) in supp (D,X), so that θ^* solves

$$\max_{\theta\in\Theta} \int \tau(r(d,\tilde{x},\ddot{x}),m(d,x,\theta))\ \mathrm{d}G(\ddot{x} \mid d,x),$$

where Θ is an appropriate finite or infinite dimensional parameter space. Stochastic identification corresponds to the existence and uniqueness of the optimizer θ^*. Clearly, structural identification is neither necessary nor sufficient for this. The problem now is to estimate θ^*.

Given a sample of n observations (Y_i, D_i, X_i), an estimator $\hat{\theta}_n$ of θ^* can be obtained as the solution to the feasible problem

$$\max_{\theta\in\Theta_n} n^{-1}\sum_{i=1}^{n} \tau(Y_i, m(D_i,X_i,\theta)),$$

where $\{\Theta_n\}$ is a suitable sequence of subsets of Θ. If $\Theta_n = \Theta$ and Θ is finite dimensional, the method is parametric. Nonparametric methods arise when Θ is infinite dimensional. For example, the method of sieves (Grenander, 1981; Chen, 2005) arises if $\{\Theta_n\}$ is increasing and $\cup_{n=1}^{\infty}\Theta$ is dense in Θ. The analysis of $\hat{\theta}_n$ as an estimator of θ^* is then standard, although the causal structure has a variety of interesting implications for estimation and inference (see White and Chalak, 2006).

An important application of this method is to estimate the conditional moments

$$\mu_k(d,x) \equiv \int \tau_k(r(d,\tilde{x},\ddot{x}))\ \mathrm{d}G(\ddot{x} \mid d,x), \qquad k = 1,2,\ldots,$$

used to construct the general conditional moment $\mu_0(d,x) = \tau_0(\mu_1(d,x),\ \mu_2(d,x),\ \ldots)$ of Theorem 4.1. With structural identification, the covariate-conditioned counterfactual moment ρ_0 equals the standard conditional moment μ_0, so to estimate ρ_0 it suffices to estimate the μ_k's. As discussed in Gourieroux, Monfort, and Trognon (1984) and White (1994, ch. 5), any conditional mean (here $E(\tau_k(Y) \mid D,X)$) can, under suitable regularity conditions, be estimated by quasi-maximum likelihood based on members of the linear exponential family, e.g., the method of weighted least squares.

Next, consider the covariate-conditioned implicit moment μ_0 of Theorem 4.3, such that

$$\int \tau(r(d,\tilde{x},\ddot{x}),\mu_0(d,x))\ \mathrm{d}G(\ddot{x}\mid d,x)=0.$$

Given structural identification, the covariate-conditioned counterfactual implicit moment ρ_0 equals μ_0. Now we parameterize the solution μ_0, specifying a parameter space Θ and a function $\theta \to m(\cdot,\cdot,\theta)$, such that $m(d,x,\theta^*)=\mu_0(d,x)$ for all (d,x) in supp (D,X), so θ^* satisfies

$$\int \tau(r(d,\tilde{x},\ddot{x}),m(d,x,\theta^*))\ \mathrm{d}G(\ddot{x}\mid d,x)=0.$$

That is, θ^* solves the implicit moment conditions $E(\tau(Y,m(D,X,\theta^*))\mid D,X)=0$. Stochastic identification holds when θ^* is the unique solution to these equations. To estimate θ^*, one can apply parametric or non-parametric versions of the methods of moments (Hansen, 1982; Ai and Chen, 2003) or empirical likelihood methods.

4.4 Weaker Conditions for Structural Identification

We now consider what happens when conditional exogeneity is not imposed. This yields weaker conditions ensuring structural identification locally, that is, at specific values (d,x). In some cases, these are also necessary and sufficient.

We first examine what happens to the relationship between the counterfactual moments ρ_k and their standard counterparts μ_k in Theorem 4.1 in the absence of B.2.

Theorem 4.6 Suppose that B.1(a) holds and let

$$s(d,x,\ddot{x})\equiv 1-dG(\ddot{x}\mid x)\,/\,dG(\ddot{x}\mid d,x)=1-dG(d\mid x)\ /\ dG(d\mid \ddot{x},x).$$

(i) Then for all $(d,\ x)\in$ supp($D,\ X$), $\int s(d,x,\ddot{x})\ dG(\ddot{x}\mid d,x))=0$; (ii) Further, let B.1(c.i) and the remaining conditions of Theorem 4.1(i) hold, and suppose that $E(s(D,X,\ddot{X})^2)<\infty$ and $E(\tau_k(Y)^2)$ $<\infty$, $k=1,2,\ldots$. Then for all $(d,x)\in$ supp (D,X), $\rho_k(d,x)$ as defined in Theorem 4.1(ii) exists and is finite, and

$$\mu_k(d,x)=\rho_k(d,x)+\gamma_k(d,x),$$

where

$$\gamma_k(d,x) \equiv \int \tau_k(r(d,\tilde{x},\ddot{x}))\ s(d,x,\ddot{x})\, dG(\ddot{x}|d,x), \qquad k = 1,2,\ldots.$$

(iii) Further, for all $(d,x) \in$ supp (D,X),

$$|\gamma_k(d,x)| \leq \sigma(d,x;\tau_k)\sigma_s(d,x), \qquad k = 1,2,\ldots,$$

where $\sigma(d,x;\tau_k) \equiv [\text{var }(\tau_k(Y) \mid (D,X) = (d,x))]^{1/2}$ and $\sigma_s(d,x) \equiv [\text{var }(s(D,X,\ddot{X}) \mid (D,X) = (d,x))]^{1/2}$. ■

The *discrepancy score* $s(d,x,\ddot{x})$ measures the relative departure from conditional independence at $(d,x,\ddot{x})$. When $s(d,x,\ddot{x}) = 0$, we have *local conditional independence at* $(d,x,\ddot{x})$. According to (i), the discrepancy score has conditional mean zero.

According to (ii), the standard conditional moment $\mu_k(d,x)$ differs from the counterfactual moment $\rho_k(d,x)$ by the *moment discrepancy* $\gamma_k(d,x)$, which, given (i), is the conditional covariance of the moment function $\tau_k(Y)$ and the discrepancy score $s(D,X,\ddot{X})$. Thus, $\gamma_k(d,x) = 0$ is the necessary and sufficient condition for structural identification of the counterfactual moment $\rho_k(d,x)$. We view this as a *local identification* result, as it is specific to a particular point (d,x). For this it suffices that $s(d,x,\ddot{x}) = 0$ for all $\ddot{x}$ in supp$(\ddot{X} \mid (D,X) = (d,x))$ Clearly, conditional exogeneity (B.2) is sufficient for this, but not necessary.

The Cauchy-Schwarz inequality gives (iii), bounding the moment discrepancy and establishing a continuity property with respect to (i) local dependence of $\tau_k(Y)$ on unobservable explanatory variables, measured by $\sigma(d,x;\tau_k)$; and (ii) local departures from conditional independence, measured by $\sigma_s(d,x)$. If either of these influences is small, then so is the moment discrepancy. Theorem 4.6(iii) thus provides a "near identification" result. The bound is the best possible in the sense that equality is attained when $|\tau_k(r(d,\tilde{x},\ddot{x}))| = |s(d,x,\ddot{x})|$ for given (d,x) and all $\ddot{x}$ in supp$(\ddot{X}|(D,X) = (d,x))$. Similar results follow by applying the Hölder inequality in place of Cauchy-Schwarz.

Recall that Theorem 4.1 treats general moments $\mu_0(d,x) = \tau_0(\mu_1(d,x),\ \mu_2(d,x),\ldots)$ and counterfactual moments $\rho_0(d,x) = \tau_0(\rho_1(d,x),\ \rho_2(d,x),\ldots)$. The *general moment discrepancy* is

$$\gamma_0(d,x) \equiv \mu_0(d,x) - \rho_0(d,x).$$

When τ_0 is an affine function of its arguments (consider the important special case of the simple covariate-conditioned average response), we have

$$\gamma_0(d, x) = \tau_0(\gamma_1(d, x), \gamma_2(d, x), \ldots).$$

It then follows immediately from Theorem 4.6(ii) that the "apparent effect" $\Delta\mu_0(d, d^*, x)$ is contaminated by the *effect discrepancy*

$$\Delta\gamma_0(d, d^*, x) = \tau_0(\Delta\gamma_1(d, d^*, x), \Delta\gamma_2(d, d^*, x), \ldots),$$

where $\Delta\gamma_k(d, d^*, x) \equiv \gamma_k(d^*, x) - \gamma_k(d, x), k = 1, 2, \ldots$.

Even when τ_0 is not affine, γ_0 depends globally and smoothly on the γ_k's, to a close approximation under plausible conditions. For brevity, we only sketch a proof. For simplicity, let τ_0 depend on $\mu \equiv (\mu_1, \ldots, \mu_k)'$ taking values in a compact set. If τ_0 is continuous, then

$$\tau_0(\mu) = \sum_{i=1}^{\infty} a_i \text{cos } (\mu'\theta_i) + \sum_{i-1}^{\infty} b_i \sin(\mu'\theta_i),$$

gives the Fourier series representation, where the a_i's and b_i's are the Fourier coefficients and the θ_i's are the appropriate multi-frequencies. It follows that

$$\tau_0(\mu) - \tau_0(\rho) = [\sum_{i=1}^{\infty} a_i \cos(\mu'\theta_i) - \sum_{i=1}^{\infty} a_i \cos(\rho'\theta_i)] + [\sum_{i=1}^{\infty} b_i \sin(\mu'\theta_i) - \sum_{i=1}^{\infty} b_i \sin(\rho'\theta_i)].$$

By straightforward trigonometric identities, we have

$$\begin{aligned} \cos(u) - \cos(v) &= 2\sin(u)\cos([v-u]/2)\sin([v-u]/2) + 2\cos(u)\sin^2([v-u]/2) \\ \sin(u) - \sin(v) &= -2\cos(u)\cos([v-u]/2)\sin([v-u]/2) + 2\sin(u)\sin^2([v-u]/2). \end{aligned}$$

Letting $\gamma \equiv \mu - \rho$ and substituting into $\tau_0(\mu) - \tau_0(\rho)$ then gives

$$\begin{aligned} \gamma_0(\mu, \gamma) &= 2\sum_{i=1}^{\infty} [a_i \sin(\mu'\theta_i) - b_i \cos(\mu'\theta_i)] \cos(\gamma'\theta_i/2)\sin(\gamma'\theta_i/2) \\ &\quad + 2\sum_{i=1}^{\infty} [a_i \cos(\mu'\theta_i) + b_i \sin(\mu'\theta_i)] \sin^2(\gamma'\theta_i/2). \end{aligned}$$

Even in this fairly general situation, the general moment discrepancy γ_0 thus depends globally and smoothly on the individual moment discrepancies γ, to a close approximation.

Theorem 4.6(iii) then ensures that the effect discrepancy $\Delta\gamma_0$ inherits a continuity property with respect to both departures from conditional independence and local dependence of the response on unobservable explanatory variables.

An important implication of this continuity is that *neglecting proxies for unobservables that have minor relevance for determining either the response or the cause of interest leads to correspondingly minor distortions in the apparent effect.* Thus, primary attention should be devoted to including proxies for the variables of most relevant for determining the response and the causes of interest.

Marginal effects are similarly affected when B.2 fails. To obtain results we add further structure, gaining analytic convenience without losing much generality. We now require that the possible values taken by $\ddot{X}$ do not depend on the realization of D, though they may depend on that of X. This is weaker than conditional independence, as the probabilities associated with these possible values can depend on the realization of D.

Recall that a σ-finite measure η is absolutely continuous with respect to a σ-finite measure ν, written $\eta \ll \nu$, if $\eta(\mathbf{B}) = 0$ for every measurable set B such that $\nu(\mathrm{B}) = 0$. If $\eta \ll \nu$, we say that ν *dominates* η and call ν a "dominating measure." The Radon-Nikodym theorem states that if $\eta \ll \nu$, then there exists a positive measurable function f $=\mathrm{d}\eta/\mathrm{d}\nu$, the *Radon-Nikodym density*, such that $\eta(\mathrm{A}) = \int_{\mathrm{A}} f \mathrm{d}\nu$ for every measurable set A. We use these notions to state the following assumption.

Assumption B.3 (a) For each $x \in \mathrm{supp}(\ X\ )$, there exists a σ-finite measure $\nu(\cdot \mid x)$ such that for each $(d, x) \in \mathrm{supp}(D, X)$, the measure $G(\mathrm{B}|\ d, x) = \int_{\mathrm{B}} \mathrm{d}G(\ddot{x} \mid d, x)$ is absolutely continuous with respect to $\nu(\cdot \mid x)$. ■

By Radon-Nikodym, there exists a conditional density, say $g(\ddot{x} \mid d, x)$, such that $\mathrm{d}G(\ddot{x} \mid d, x) = g(\ddot{x} \mid d, x)\ \mathrm{d}\nu(\ddot{x} \mid x)$. Conditional dependence arises whenever $g(\ddot{x} \mid d, x)$ depends non-trivially on d. Next, we impose differentiability and domination conditions on g, using the convention specified following A.3.

Assumption B.4 For given j, $(\partial g/\partial d_j)(\ddot{x} \mid d, x)$ is dominated on C_j by a function integrable with respect to $\nu(\cdot \mid x)$ at (d, x). ■

We also impose the following analog of A.3:

Assumption B.5 For given j and $k = 1, 2, \ldots$, $(\partial((\tau_k \circ r)g) \;/\; \partial d_j)(d, x, \ddot{x})$ is dominated on C_j by a function integrable with respect to $\nu(\cdot \mid x)$ at (d, x). ■

The differentiability of g implicit in B.4 and differentiability of $\tau_k \circ r$ with respect to d_j implicit in B.5 ensure existence of the product derivative $\partial((\tau_k \circ r)g)/\partial d_j$.

Theorem 4.7 Suppose Assumptions B.1(a), B.3(a), and B.4 hold. (i) Then

$$\int (\partial \log g/\partial d_j)(\ddot{x} \mid d, x)\; g(\ddot{x} \mid d, x)\; \mathrm{d}\nu(\ddot{x} \mid x) = 0;$$

(ii) Further, let B.1(c.i), B.5, and the remaining conditions of Theorem 4.1 hold. Then the functions $d_j \to \mu_k(d, x)$, $k = 1, 2, \ldots$, are differentiable on C_j, and

$$\begin{aligned}(\partial\mu_k/\partial d_j)(d, x) &= \int (\partial(\tau_k \circ r)/\partial d_j)(d, \tilde{x}, \ddot{x})\; g(\ddot{x} \mid d, x)\; \mathrm{d}\nu(\ddot{x} \mid x) \\ &\quad + \int \tau_k(r(d, \tilde{x}, \ddot{x}))(\partial \log g/\partial d_j)(\ddot{x} \mid d, x)\; g(\ddot{x} \mid d, x)\; \mathrm{d}\nu(\ddot{x} \mid x) \\ &\equiv \xi_{k,j}(d, x) + \delta_{k,j}(d, x).\end{aligned}$$

(iii) If, in addition, $E([(\partial \log g/\partial d_j)(\ddot{X} \mid D, X)]^2) < \infty$ and $E(\tau_k(Y)^2) < \infty$, $k = 1, 2, \ldots$, then

$$|\delta_{k,j}(d, x)| \leq \sigma(d, x; \tau_k)\sigma_j(d, x),$$

where $\sigma_j(d, x) \equiv [\int \{(\partial \log g/\partial d_j)(\ddot{x} \mid d, x)\}^2\; g(\ddot{x} \mid d, x)\; \mathrm{d}\nu(\ddot{x} \mid x)]^{1/2}$. ■

The *marginal discrepancy scores* $(\partial \log g/\partial d_j)(\ddot{x} \mid d, x)$ quantify local departures from conditional exogeneity. Result (i) states that these scores have conditional mean zero. With conditional exogeneity, they are identically zero.

Result (ii) decomposes the "apparent marginal moment effect" $(\partial\mu_k/\partial d_j)(\ddot{x} \mid d, x)$ into two pieces: the true marginal moment effect $\xi_{k,j}(d, x)$ and the *marginal moment effect discrepancy* $\delta_{k,j}(d, x)$. It follows that the marginal moment effect is structurally identified (locally at (d, x)) if and only if $\delta_{k,j}(d, x) = 0$. Conditional exogeneity is sufficient but not necessary for this.

Result (iii) bounds the magnitude of the marginal moment effect discrepancy: this is jointly controlled by local dependence of the response on unobservable explanatory variables, measured by $\sigma(d, x; \tau_k)$, and by local departures from conditional exogeneity, measured by $\sigma_j(d, x)$. Parallel to

Theorem 4.6(iii), this near identification result ensures a continuity property: neglecting proxies for unobservables that have minor relevance for the determination of either the response or the causes of interest leads to correspondingly minor distortions in the apparent marginal moment effects.

Also as in the case of Theorem 4.6, the bound in Theorem 4.7 is the best possible, as equality holds when $|\tau_k(r(d,\tilde{x},\ddot{x}))| = |(\partial \log g/\partial d_j)\ (\ddot{x} \mid d,x)|$ for given (d,x) and all $\ddot{x}$ in $\text{supp}(\ddot{X} \mid (D,X) = (d,x))$. Similar results hold by applying the Hölder inequality in place of Cauchy-Schwarz.

Results for the relation between the derivatives of the general moments $\mu_0(d,x)$ and $\rho_0(d,x)$ follow straightforwardly. For brevity, we sketch their derivation under simplifying assumptions. Again, let τ_0 depend on $\mu \equiv (\mu_1,\ldots,\mu_k)'$. The chain rule gives

$$\partial\mu_0/\partial d_j = \nabla'\tau_0(\mu)(\partial\mu/\partial d_j), \qquad \partial\rho_0/\partial d_j = \nabla'\tau_0(\rho)(\partial\rho/\partial d_j),$$

where $\nabla\tau_0$ is the $k \times 1$ gradient vector of τ_0 with respect to its arguments, and $\partial\mu/\partial d_j$ and $\partial\rho/\partial d_j$ are the $k \times 1$ vectors containing the partial derivatives with respect to d_j of the elements of μ and ρ respectively. We suppress the dependence of these expressions on (d,x). Adding and subtracting appropriately gives

$$\partial\mu_0/\partial d_j - \partial\rho_0/\partial d_j = \nabla'\tau_0(\mu)[(\partial\mu/\partial d_j) - (\partial\rho/\partial d_j)]$$

$$+[\nabla'\tau_0(\mu) - \nabla'\tau_0(\rho)](\partial\mu/\partial d_j) - [\nabla'\tau_0(\mu) - \nabla'\tau_0(\rho)][(\partial\mu/\partial d_j) - (\partial\rho/\partial d_j)]$$

$$= \nabla'\tau_0(\mu)\delta_j + \nabla'\gamma_0(\mu,\gamma)(\partial\mu/\partial d_j) - \nabla'\gamma_0(\mu,\gamma)\delta_j,$$

$$\equiv \delta_0(\mu,\gamma,\delta_j)$$

where δ_j is the vector whose elements are the marginal moment effect discrepancies and $\nabla'\gamma_0(\mu,\gamma)$ $= \nabla'\tau_0(\mu) - \nabla'\tau_0(\rho)$ holds with smoothness assumptions on τ_0 sufficient to ensure that the Fourier series approximation used above holds with respect to a suitable Sobolev norm. The *marginal general moment effect discrepancy* $\delta_0(\mu,\gamma,\delta_j)$ thus depends on both the moment discrepancy vector γ and the marginal moment discrepancy δ_j. This can vanish for specific values of (d,x) under special circumstances. It vanishes for all (d,x) under conditional exogeneity.

We subsume study of optimizer-based distributional aspect discrepancies into the study of implicit moment discrepancies, viewing the optimizer-based distributional aspects as implicit moments defined by the first order conditions of the underlying optimization. Accordingly, we omit explicit complements to Theorem 4.2 analogous to those provided by Theorems 4.6 and 4.7 for Theorem 4.1.

Significant challenges to obtaining a tractable representation for the general implicit moment discrepancy $\gamma_0(d,x) \equiv \mu_0(d,x) - \rho_0(d,x)$ arise from the implicit nonlinear definitions of μ_0 and ρ_0. Accordingly, to the extent that such implicitly defined moments cannot be well approximated by the explicit moments analyzed in Theorems 4.1 and 4.6, we leave their analysis aside here. On the other hand, an analysis of marginal effect discrepancies analogous to Theorem 4.7 is straightforward.

To succinctly state our next result, we now write

$$\int \tau_\mu \; g_d \; \mathrm{d}\nu = \int \tau(r(d,\tilde{x},\ddot{x}), \mu_0(d,x)) \; g(\ddot{x} \mid d,x) \; \mathrm{d}\nu(\ddot{x} \mid x).$$

We elaborate Assumption B.3(a) by imposing

Assumption B.3(b) For each $x \in \operatorname{supp}(X)$, the measure defined by $G(B|x) = \int_B \mathrm{d}G(\ddot{x} \mid x)$ is absolutely continuous with respect to $\nu(\cdot \mid x)$. ■

By Radon-Nikodym, there exists $g(\ddot{x} \mid x)$ such that $\mathrm{d}G(\ddot{x} \mid x) = g(\ddot{x} \mid x) \; \mathrm{d}\nu(\ddot{x} \mid x)$. When the integral of Theorem 4.3(iii) exists, B.3(b) allows us to write

$$\int \tau_\rho \; g \; \mathrm{d}\nu = \int \tau(r(d,\tilde{x},\ddot{x}), \rho_0(d,x;\tau)) \; g(\ddot{x} \mid x) \; \mathrm{d}\nu(\ddot{x} \mid x).$$

When the referenced derivatives exist, we now write $\partial\mu_0/\partial d_j$ as the $\lambda \times 1$ vector containing the derivatives $(\partial/\partial d_j)\mu_{0,i}(d,x)$, $i = 1, \ldots, \lambda$, and $\partial\rho_0/\partial d_j$ is now the $\lambda \times 1$ vector containing $(\partial/\partial d_j)\rho_{0,i}(d,x), i = 1, \ldots, \lambda$. For $\tau(r,m)$, let $\nabla_r \tau_\rho$ denote the $\lambda \times 1$ vector containing $(\partial/\partial r)\tau_i(r(d,\tilde{x},\ddot{x}),$ $\rho_0(d,x)), i = 1, \ldots, \lambda$; let $\nabla_r \tau_\mu$ denote the $\lambda \times 1$ vector containing $(\partial/\partial r)\tau_i(r(d,\tilde{x},\ddot{x}),\ \mu_0(d,x)), i = 1, \ldots, \lambda$; let $\nabla'_m \tau_\mu$ denote the $\lambda \times \lambda$ matrix whose ith row has elements $(\partial/\partial m_j)\tau_i(r(d,\tilde{x},\ddot{x}),\ \mu_0(d,x))$, $j = 1, \ldots, \lambda$, $i = 1, \ldots, \lambda$; and let $\nabla'_m \tau_\rho$ denote the $\lambda \times \lambda$ matrix whose ith row has elements $(\partial/\partial m_j)\tau_i(r(d,\tilde{x},\ddot{x}),\ \rho_0(d,x))$, $j = 1, \ldots, \lambda$, $i = 1, \ldots, \lambda$. When the integrals and inverses exists, we also define $D_\mu \equiv -[\int \nabla'_m \tau_\mu \; g_d \; d\nu]^{-1}$, $D_\rho \equiv -[\int \nabla'_m \tau_\rho \; g \; d\nu]^{-1}$.

Assumption B.6 (a) The elements of $\tau(r(d,\tilde{x},\ddot{x}),\ \mu_0(d,x)) \; g(\ddot{x} \mid d,x)$ are dominated on C_j by a function integrable with respect to $\nu(\cdot \mid x)$ at (d,x). (b) The elements of $\tau(r(d,\tilde{x},\ddot{x}),\ \rho_0(d,x))$ are dominated on C_j by a function integrable with respect to $\mathrm{d}G(\cdot \mid x)$ at (d,x). ■

Theorem 4.8. (i) Suppose the conditions of Theorem 4.3(i) and B.1(c.i) hold, that $E(s(D,X,\ddot{X})^2)$ $< \infty$, and that $E(\tau(Y,m)^2) < \infty$ for each $m \in M \subset \mathbb{R}^\lambda$. Then for each (d,x,m) in supp $(D,X) \times$

M the conditional expectation

$$\varphi_\tau(d,x,m) = \int \tau(r(d,\tilde{x},\ddot{x}),m)\ \mathrm{d}G(\ddot{x} \mid x)$$

exists and is finite. (ii) Further, let τ, r, and $x \to G(\cdot \mid x)$ be such that for each (d,x,m) in supp $(D,X) \times M$, φ_τ is differentiable on a neighborhood of (d,x,m), the $\lambda \times \lambda$ matrix $\nabla_m \varphi_\tau(d,x,m)$ is non-singular, and $\varphi_\tau(d,x,m) = 0$. Then there exists a unique function ρ_0 such that for each $(d,x) \in$ supp (D,X), ρ_0 is differentiable at (d,x) and

$$\int \tau(r(d,\tilde{x},\ddot{x}),\ \rho_0(d,x))\ \mathrm{d}G(\ddot{x} \mid x) = 0.$$

(iii) If B.3(b) and B.6(b) also hold, if τ is differentiable and $(\partial r/\partial d_j)(d,\tilde{x},\ddot{x})$ exists on $C_j \times$ supp $(\ddot{X} \mid \tilde{x})$, and if $\int \nabla'_m \tau_\rho\ g\ \mathrm{d}\nu$ exists and is finite and non-singular, then

$$\partial\rho_0/\partial d_j = D_\rho \int \nabla_r \tau_\rho\ (\partial r/\partial d_j)\ g\ \mathrm{d}\nu.$$

(iv) If in addition B.3(c) and B.6(c) hold, then $\partial\mu_0/\partial d_j = \partial\rho_0/\partial d_j + \delta_{0,j}$, where

$$\begin{aligned}
\delta_{0,j} &= D_\mu \int \tau_\mu\ (\partial \log g_d/\partial d_j)\ g_d\ \mathrm{d}\nu + \int [D_\mu \nabla_r \tau_\mu - D_\rho \nabla_r \tau_\rho\ g/g_d]\ (\partial r/\partial d_j)\ g_d\ \mathrm{d}\nu \\
&= D_\mu \int \tau_\mu\ (\partial \log g_d/\partial d_j)\ g_d\ \mathrm{d}\nu + D_\mu \int \nabla_r \tau_\mu\ (1 - g/g_d)\ (\partial r/\partial d_j)\ g_d\ \mathrm{d}\nu \\
&\quad + \int (D_\mu \nabla_r \tau_\mu - D_\rho \nabla_r \tau_\rho)\ (1 - g/g_d)\ (\partial r/\partial d_j)\ g_d\ \mathrm{d}\nu \\
&\quad + \int (D_\mu \nabla_r \tau_\mu - D_\rho \nabla_r \tau_\rho)\ (\partial r/\partial d_j)\ g_d\ \mathrm{d}\nu.
\end{aligned}$$

(v) Further,

$$|\delta_{0,j}(d,x)| \leq \sigma(d,x;\tau_\mu)\sigma_j(d,x) + \sigma_{\mu,\rho}(d,x)\sigma_{r,j}(d,x),$$

where $\sigma(d,x;\tau_\mu) \equiv [\text{var}\ (\tau(Y,\mu_0(D,X) \mid (D,X) = (d,x)]^{1/2}$, $\sigma_j(d,x)$ is as previously defined, $\sigma^2_{\mu,\rho}(d,x) \equiv \int [D_\mu \nabla_r \tau_\mu - D_\rho \nabla_r \tau_\rho g/g_d]^2 g_d\ \mathrm{d}\nu$, and $\sigma^2_{r,j}(d,x) \equiv \int (\partial r/\partial d_j)^2 g_d\ \mathrm{d}\nu$. ■

The functions appearing in (iii) and (iv) above are implicitly evaluated at (d,x) as specified in B.6.

To gain some insight into the *marginal implicit moment effect discrepancy* $\delta_{0,j}$ of (iv), we first observe that each of its components vanishes under conditional exogeneity, as then $(\partial \log g_d/\partial d_j) = 0$, $(1 - g/g_d) = 0$, and $D_\mu \nabla_r \tau_\mu$ - $D_\rho \nabla_r \tau_\rho = 0$ (by Theorem 4.3). When conditional exogeneity fails but the true marginal effect $(\partial r/\partial d_j)(d,\tilde{x},\ddot{x})$ is zero for all $\ddot{x}$ in supp $(\ddot{X} \mid (D,X)=(d,x))$, the true marginal implicit moment effect vanishes ($(\partial\rho_0/\partial d_j)(d,x)=0$), but the apparent effect becomes

$$\delta_{0,j} \equiv D_\mu \int \tau_\mu \ (\partial \log g_d/\partial d_j) \ g_d \ \mathrm{d}\nu.$$

From Theorem 4.7(i) $\delta_{0,j}$ is thus a linear combination of conditional covariances between τ_μ and $\partial \log g_d/\partial d_j$.

The simple but important special case $\tau(r,m)=r-m$ offers further insight. Here $D_\mu = D_\rho = 1$ and $\nabla_r \tau_\mu = \nabla_r \tau_\rho = 1$. In this case, we have

$$\delta_{0,j} \equiv \int \tau_\mu(\partial \log g_d/\partial d_j) \ g_d \ \mathrm{d}\nu + \int (1-g/g_d) \ (\partial r/\partial d_j) \ g_d \ \mathrm{d}\nu.$$

(Compare this with the result of Theorem 4.7(ii).) The first term is the conditional covariance between τ_μ and $(\partial \log g_d/\partial d_j)$, and the second term is that between $(1-g/g_d)$ and $(\partial r/\partial d_j)$. Similar results hold with $\tau\,(r,m)=\tau_k(r)-m$.

This result affords considerable opportunity for exploration of special cases of interest (for example, when $\lambda = 1$, consider the anti-symmetric case in which $\tau(r,m) = -\tau(m,r)$, which applies to the conditional median). For brevity, we leave exploration of these cases to other work.

The first three components of $\delta_{0,j}$ in (iv) are clearly interpretable as conditional covariances, given that $(\partial \log g_d/\partial d_j)$ and $(1-g/g_d)$ have conditional mean zero. The fourth term is not obviously a covariance because there is no need for $(\partial r/\partial d_j)$ to have conditional mean zero and it is not obvious that $D_\mu \nabla_r \tau_\mu$ - $D_\rho \nabla_r \tau_\rho$ has conditional mean zero. Similarly, the terms $\sigma^2_{\mu,\rho}$ and $\sigma^2_{r,j}$ appearing in (v) are not necessarily variances; instead, they measure departures from zero.

Result (v) provides a continuity result, generalizing that of Theorem 4.7(iii).

5 Unconfoundedness and Conditional Exogeneity

Unconfoundedness is a standard notion in the treatment effects literature. This notion was originally introduced by Rosenbaum and Rubin (1983) for a single binary treatment, D, as "ignorable treatment assignment;" it requires

$$(Y(0), Y(1)) \perp D \mid X,$$

where $Y(0)$ is the response in the absence of treatment ($D = 0$), $Y(1)$ is the response in the presence of treatment ($D = 1$), and X represents covariates. As Imbens (2004, p.7) notes, this condition has also been called the "conditional independence condition" by Lechner (1999, 2001), and "selection on observables" (Heckman and Robb, 1985), following work of Barnow, Cain, and Goldberger (1980).

In the treatment effects litereature, the response function generating Y(0) and Y(1) is typically not explicitly specified. In contrast, the settable systems framework explicitly imposes structure provided by economic theory to represent the response as an unknown function r of causes of interest, D, and relevant ancillary causes, Z, as modified by attributes, A. With a single binary treatment, we write

$$Y(0) = r(0, Z, A), \qquad Y(1) = r(1, Z, A).$$

Analogous to White (2005a, Proposition 3.2), Assumptions B.1 and B.2 imply $(Y(0), Y(1)) \perp D \mid X$.

Hirano and Imbens (2004) extend the unconfoundedness concept to the case of a continuous treatment D taking values in a set $\boldsymbol{D}$. Their "weak unconfoundedness" condition requires

$$Y(d) \perp D \mid X \text{ for all } d \text{ in } \boldsymbol{D},$$

where $Y(d)$ represents the "potential" or "counterfactual" response to treatment d. In fact, their approach extends to multiple treatments where components may be binary, categorical, or continuous, as permitted here. In this context, $\boldsymbol{D} = \text{supp}(D)$. The generalized propensity score is again the conditional density of D given X.

As is standard in the treatment effects literature, Hirano and Imbens (2004) do not specify the response function generating $Y(d)$. Nevertheless, by taking $Y(d) = r(d, Z, A)$, it is straightforward to show that B.1 and conditional exogeneity imply weak unconfoundedness.

By not specifying a response function, one operates in a general context, as there is no need to commit to any theory about how the response is determined, apart from its potential dependence on the causes of interest. The price paid for this agnosticism in non-experimental contexts is that one has correspondingly little guidance as to the construction of proper covariates. Moreover, in the absence of randomization, the availability of covariates is generally necessary for the identification of causal effects. Even if one desired to proceed atheoretically, this would fail in practice, as economic theory must inevitably play some role in the selection of covariates.

In contrast, by allowing economic theory to play its full and natural role in specifying the variables potentially determining the response and the relations between these and other variables of the economic system, one gains broad insight into the selection of legitimate and illegitimate covariates, as the discussion of Section 4.2 demonstrates. The settable system approach provides a framework in which economic theory can operate, designed specifically to be compatible with the higher-level atheoretic structures of the treatment effects framework. Once such theory-based structure (Assumption B.1) is in place, the higher-level unconfoundedness conditions of the treatment effect literature can operate in the background, as ensured by conditional exogeneity.

The artificial intelligence DAG approach to causal identification also exploits conditional independence conditions to identify causal effects (e.g., Verma and Pearl, 1991). Nevertheless, we prefer to refer to B.2 as a conditional exogeneity rather than as simply a conditional independence condition, as conditional exogeneity references causes of interest, unobserved explanatory variables, and covariates, whereas without further elaboration conditional independence does not.

6 Summary and Conclusion

This paper unifies three complementary approaches to defining and identifying causal effects: the classical structural equations approach of the Cowles Commision; the treatment effects framework of Rubin (1974) and Rosenbaum and Rubin (1983); and the Directed Acyclic Graph (DAG) approach of Pearl. The settable system framework described here not only nests each of these prior approaches, but also affords significant advantages and improvements relative to each. Among other things, the settable system approach permits identification and estimation of causal effects without requiring exogenous instruments, generalizing the classical structural equations approach; it relaxes the stable unit treatment value assumption of the treatment effect approach; and it accommodates mutual causality and attributes, generalizing the DAG approach.

This work thus complements and creates the opportunity for further extensions of the work of Heckman, Ichimura, and Todd (1997, 1998), Angrist (1998), Hahn (1998), Hirano and Imbens (2001), Hirano, Imbens, and Ridder (2003), Imbens and Newey (2003), Matzkin (2003, 2004, 2005), Heckman, Urzua, and Vytlacil (2005), and Heckman and Vytlacil (2005), among others.

The settable system framework consists of an underlying stochastic structure, together with a causal structure that rests on this stochastic foundation. It permits precise definitions of cause and effect and of interventions and counterfactuals, and it identifies the objects of experimental control. It enables straightforward provision of conditions for structural identification of a broad range of causal effects, based on the equality of counterfactual objects with stochastically meaningful objects. The settable system framework yields extensive guidance, not previously available, as to the construction of suitable and unsuitable covariates, revealing the crucial roles played by economic theory and domain knowledge in this construction.

Given structural identification, it follows immediately that these effects can be informatively estimated using standard estimates of their stochastically meaningful counterparts. We pursue this in White and Chalak (2006), where we also study tests for conditional exogeneity.

Our focus here is on recursive settable systems. Analysis of other partitioned settable systems should yield identification results for effects associated with mutually causal relationships, direct or indirect. Versions of instrumental variables methods may yield effective estimators of such effects.

A related direction for further research is to study identification and estimation in the absence of the requirement that the causes of interest do not cause the ancillary causes (Assumption B.1(c.i)). This raises the possibility of decomposing total effects into direct and indirect components. Chalak and White (2006) constitutes a beginning for this inquiry, revealing further extensions of the concept of instrumental variables. Also of interest is to construct statistical tests for validity of B.1(c.i).

Finally, it is our hope that empirical application of the settable system framework and the reinterpretation of standard methods that it affords will enable clearer and more robust understanding of causal effects in economic science.

7 Mathematical Appendix

Proof of Proposition 3.1 (i) Given A.1(a, b) and $E(Y) < \infty$, $E(Y \mid D, \tilde{X})$ exists and is finite by Billingsley (1979, p.395). (ii) Apply theorem 34.5 of Billingsley (1979). ■

Proof of Theorem 3.2 (i) Given A.1(a, b) and $E(Y) < \infty$, Proposition 3.1 gives $\tilde{\mu}(d, \tilde{x}) = \int r(d, \tilde{x}, \ddot{x})\, d\tilde{G}(\ddot{x} \mid d, \tilde{x})$. Assumption A.2 then implies $\int r(d, \tilde{x}, \ddot{x}) d\tilde{G}(\ddot{x} \mid d, \tilde{x}) = \int r(d, \tilde{x}, \ddot{x}) d\tilde{G}(\ddot{x} \mid \tilde{x})$ $= \tilde{\rho}(d, \tilde{x})$, ensuring both existence of $\tilde{\rho}(d, \tilde{x})$ and $\tilde{\rho}(d, \tilde{x}) = \tilde{\mu}(d, \tilde{x})$. (ii) Assumption A.3 permits application of Bartle (1966, corollary 5.9) establishing the differentiability of $\tilde{\rho}$ and $\tilde{\mu}$ (by (i)) and ensuring the validity of an interchange of derivative and integral, giving

$$\begin{aligned}(\partial\mu/\partial d_j)(d,\tilde{x}) &= (\partial\tilde{\rho}/\partial d_j)(d,\tilde{x}) \\ &= \int(\partial r/\partial d_j)(d,\tilde{x},\ddot{x})d\tilde{G}(\ddot{x}\mid\tilde{x}) \\ &= \int(\partial r/\partial d_j)(d,\tilde{x},\ddot{x})d\tilde{G}(\ddot{x}\mid d,\tilde{x}) \\ &\equiv \tilde{\xi}_j(d,\tilde{x}),\end{aligned}$$

where (i) gives the first equality, A.3 ensures the interchange in the second, and A.1.(c) ensures the absence of terms involving $(\partial z/\partial d_j)$ in the second. The third equality follows by A.2. ■

Proof of Proposition 3.3 $\tilde{X}$ and $\ddot{X}$ exist by Assumption A.1(a). Given $\ddot{X}\perp U\mid\tilde{X}$, it follows from Dawid (1979, lemma 4. 1) that $(\tilde{X},\ddot{X})\perp(U,\tilde{X})\mid\tilde{X}$. Applying Dawid (1979, lemma 4.2(i)) twice gives $\ddot{X}\perp c(\tilde{X},U)\mid\tilde{X}$. ■

Proof of Corollary 3.4 Immediate from Proposition 3.3 and Theorem 3.2. ■

Proof of Proposition 4.1. (i) For given $k=1,2,\ldots$, the proof is identical to that of Proposition 3.1(i), *mutatis mutandis* (replacing A.1(a, b) with B.1(a, b), $\tilde{X}$ with X, r with $\tau_k\circ r$, $\tilde{\mu}$ with μ_k, $\tilde{H}$ with H, and $\tilde{G}$ with G). The measurability of μ_0 follows by measurabiity of compositions of measurable functions. (ii) For given $k=1,2,\ldots$, the proof is identical to that of Proposition 3.2(i), *mutatis mutandis*. Measurability follows for ρ_0 just as it does for μ_0. That $\rho_0=\mu_0$ follows immediately from $\rho_k=\mu_k$, $k=1,2,\ldots$. ■

Proof of Theorem 4.2. (i) Given $E(\tau(Y,m))<\infty$, the existence and finiteness of $\varphi_{\tau,d}\ (d,x,m)$ follow from Billingsley (1979, p.395). (ii) For each (d,x) in supp (D,X), the existence of the non-empty, compact-valued, upper hemi-continuous correspondence $\mu_0(d,x)$ follows from the Theorem of the Maximum (Berge, 1963) under the stated conditions. (iii) If B.1(c.i) and B.2 also hold, then for each $(d,x,m)\in(D,X)\times\mathbb{R}^\lambda\ \varphi_{\tau,d}\ (d,x,m)=\varphi_\tau\ (d,x,m)$. Setting $\rho_0\ (d,x)=\mu_0\ (d,x)$ completes the proof. ■

Proof of Theorem 4.3. (i) Given $E(\tau(Y,m))<\infty$, the existence and finiteness of $\varphi_{\tau,d}\ (d,x,m)$ follow from Billingsley (1979, p. 395). (ii) The existence, uniqueness, and differentiability of μ_0 follow immediately under the given conditions from the implicit function theorem (e.g., Chiang, 1984, pp. 210-211). (iii) If B.1(c.i) and B.2 also hold, then for each $(d,x)\in$ supp (D,X)

$$\int\tau(r(d,\tilde{x},\ddot{x}),\mu_0(d,x))\ \mathrm{d}G(\ddot{x}\mid d,x)=\int\tau(r(d,\tilde{x},\ddot{x}),\mu_0(d,x))\ \mathrm{d}G(\ddot{x}\mid x).$$

Setting $\rho_0(d, x) = \mu_0(d, x)$ completes the proof. ■

Proof of Theorem 4.4. Identical to that of Proposition 3.3, *mutatis mutandis*. ■

Proof of Corollary 4.5. Immediate from Proposition 4.4 and Theorem 4.1, 4.2, or 4.3. ■

Proof of Theorem 4.6. Given B.1(a), the densities $\mathrm{d}G(\ddot{x} \mid d, x)$ and $\mathrm{d}G(\ddot{x} \mid x)$ exist and can be written $\mathrm{d}G(\ddot{x} \mid d, x) = \mathrm{d}G(d, x, \ddot{x}) \ / \ \mathrm{d}G(d, x)$ and $\mathrm{d}G(\ddot{x} \mid x) = \mathrm{d}G(x, \ddot{x}) \ / \ \mathrm{d}G(x)$. Consequently, $\mathrm{d}G(\ddot{x} \mid x) \ / \ \mathrm{d}G(\ddot{x} \mid d, x) = [\mathrm{d}G(x, \ddot{x}) \ / \ \mathrm{d}G(x)] \ / \ [\mathrm{d}G(d, x, \ddot{x}) \ / \ \mathrm{d}G(d, x)] = [\mathrm{d}G(d, x) \ / \ \mathrm{d}G(x)] \ / \ [\mathrm{d}G (d, x, \ddot{x}) \ / \ \mathrm{d}G(x, \ddot{x})] = \mathrm{d}G(d \mid x) \ / \ \mathrm{d}G(d \mid \ddot{x}, x)$. (i) We have $\int s(d, x, \ddot{x}) \ \mathrm{d}G(\ddot{x} \mid d, x) = \int \{[\mathrm{d}G(\ddot{x} \mid d, x)$ - $\mathrm{d}G(\ddot{x} \mid x)] \ / \ \mathrm{d}G(\ddot{x} \mid d, x)\} \ \mathrm{d}G(\ddot{x} \mid d, x) = \int [\mathrm{d}G(\ddot{x} \mid d, x) - \mathrm{d}G(\ddot{x} \mid x)] = 0$, given that $\mathrm{d}G(\ddot{x} \mid d, x)$ and $\mathrm{d}G(\ddot{x} \mid x)$ are each conditional densities. (ii) Given B.1(a, b) and the conditions on τ_k, Proposition 4.1(i) gives $\mu_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x})) \ \mathrm{d}G(\ddot{x} \mid d, x)$. Adding and subtracting appropriately, we have

$$\begin{aligned}
\rho_k(d, x) \quad &\equiv \quad \int \tau_k(r(d, \tilde{x}, \ddot{x})) \ \mathrm{d}G(\ddot{x} \mid x) \\
&= \quad \int \tau_k(r(d, \tilde{x}, \ddot{x})) \ \mathrm{d}G(\ddot{x} \mid d, x) + \int \tau_k(r(d, \tilde{x}, \ddot{x})) \ [\mathrm{d}G(\ddot{x} \mid x) - \ \mathrm{d}G(\ddot{x} \mid d, x)] \\
&= \quad \mu_k(d, x) + \int \tau_k(r(d, \tilde{x}, \ddot{x}))\{[\mathrm{d}G(\ddot{x}|x) - \ \mathrm{d}G(\ddot{x} \mid d, x)] \ / \ \mathrm{d}G(\ddot{x} \mid d, x)\} \ \mathrm{d}G(\ddot{x} \mid d, x) \\
&= \quad \mu_k(d, x) - \int \tau_k(r(d, \tilde{x}, \ddot{x})) \ s(d, x, \ddot{x}) \ \mathrm{d}G(\ddot{x} \mid d, x) \\
&= \quad \mu_k(d, x) - \gamma_k(d, x).
\end{aligned}$$

The existence of $\mu_k(d, x)$ follows given $E(\tau_k(Y)) < \infty$ and the existence of $\gamma_k(d, x)$ follows from the imposed second moment conditions and the Cauchy-Schwarz inequality. It follows that $\rho_k(d, x)$ exists and $\mu_k(d, x) = \rho_k(d, x) + \gamma_k(d, x)$. (iii) The result follows immediately from the Cauchy-Schwartz inequality, applied to (ii) and using (i). ■

Proof of Theorem 4.7. (i) Assumptions B.1(a) and B.3(a) ensure that for each (d, x) in supp (D, X), $g(\ddot{x} \mid d, x) = \mathrm{d}G(\ddot{x} \mid d, x) \ / \ \mathrm{d}\nu(\ddot{x} \mid x)$ is a density by the Radon-Nikodym theorem (e.g., Bartle, 1966, theorem 8.9), so $\int g(\ddot{x} \mid d, x) \ \mathrm{d}\nu(\ddot{x} \mid x) = 1$. Assumption B.4 ensures that the left hand expression above is differentiable with respect to d_j by Bartle (1966, corollary 5.9). Differentiating both sides of this equality with respect to d_j gives

$$(\partial/\partial d_j) \int g(\ddot{x} \mid d, x) \ \mathrm{d}\nu(\ddot{x} \mid x) = 0.$$

Assumption B.4 further justifies interchanging the derivative and integral on the left by Bartle (1966, corollary 5.9), so that

$$\int (\partial g/\partial d_j)(\ddot{x} \mid d, x) \; \mathrm{d}\nu(\ddot{x} \mid x) = 0.$$

Substituting $(\partial g/\partial d_j)(\ddot{x} \mid d, x) = (\partial \log g/\partial d_j)(\ddot{x} \mid d, x) \; g(\ddot{x} \mid d, x)$ delivers the desired result. (ii) Given B.1(a, b) and the conditions on τ_k, Proposition 4.1(i) gives $\mu_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x})) \; \mathrm{d}G(\ddot{x} \mid d, x)$. Substituting $\mathrm{d}G(\ddot{x} \mid d, x) = g(\ddot{x} \mid d, x) \; \mathrm{d}\nu(\ddot{x} \mid x)$ gives

$$\mu_k(d, x) = \int \tau_k(r(d, \tilde{x}, \ddot{x})) \; g(\ddot{x} \mid d, x) \; \mathrm{d}\nu(\ddot{x} \mid x).$$

Assumption B.5 permits application of Bartle (1966, corollary 5.9) establishing the differentiability of μ_k and ensuring the validity of an interchange of derivative and integral, giving

$$\begin{aligned} (\partial \mu_k/\partial d_j)(d, x) &= \int (\partial((\tau_k \circ r)g)/\partial d_j)(d, x, \ddot{x}) \; \mathrm{d}\nu(\ddot{x} \mid x) \\ &= \int (\partial(\tau_k \circ r)/\partial d_j)(d, \tilde{x}, \ddot{x}) \; g(\ddot{x} \mid d, x) \; \mathrm{d}\nu(\ddot{x} \mid x) \\ &\quad + \int \tau_k(r(d, \tilde{x}, \ddot{x})) \; (\partial g/\partial d_j)(\ddot{x} \mid d, x) \; \mathrm{d}\nu(\ddot{x} \mid x), \end{aligned}$$

where B.1(c.i) ensures the absence of terms involving $(\partial x/\partial d_j)$ in the second equality. Substituting $(\partial g/\partial d_j)(\ddot{x} \mid d, x) = (\partial \log g/\partial d_j)(\ddot{x} \mid d, x) \; g(\ddot{x} \mid d, x)$ delivers the desired result. (iii) The result follows immediately from the Cauchy-Schwarz inequality, applied to (ii) and using (i). ■

Proof of Theorem 4.8 (i) We write

$$\begin{aligned} \varphi_\tau(d, x, m) &\equiv \int \tau(r(d, \tilde{x}, \ddot{x}), m) \; \mathrm{d}G(\ddot{x} \mid x) \\ &= \int \tau(r(d, \tilde{x}, \ddot{x}), m)[\mathrm{d}G(\ddot{x} \mid x) \; / \; \mathrm{d}G(\ddot{x} \mid d, x)] \; \mathrm{d}G(\ddot{x} \mid d, x). \end{aligned}$$

The imposed second moment conditions ensure the existence and finiteness of this integral by Cauchy-Schwarz. (ii) The existence, uniqueness, and differentiability of ρ_0 follow immediately under the given conditions from the implicit function theorem (e.g., Chiang, 1984, pp. 210-211). (iii) Given the assumed differentiability of φ_τ and Assumption B.6(b), we have

$$(\partial/\partial d_j)\varphi_\tau(d, x, \rho_0(d, x)) = \int (\partial/\partial d_j)\tau(r(d, \tilde{x}, \ddot{x}), \rho_0(d, x,)) \; \mathrm{d}G(\ddot{x} \mid x) = 0,$$

where the interchange of integral and derivative is justified by Bartle (1966, corollary 5.9), and the equality holds because $\varphi_\tau(d, x, \rho_0(d, x)) = 0$ for all (d, x) in supp (D, X). Using the assumed differentiability of τ and r, the differentiability of ρ_0 ensured by (ii), and the chain rule gives

$$\int [\nabla_r \tau_\rho (\partial r / \partial d_j) + \nabla'_m \tau_\rho \ (\partial \rho_0 / \partial d_j)] \ \mathrm{d}G(\ddot{x} \mid x) = 0,$$

where we exploit the notation introduced preceding Theorem 4.8 in the text. Solving for $\partial \rho_0 / \partial d_j$ given the assumed existence of $D_\rho \equiv -[\int \nabla'_m \tau_\rho \ \mathrm{d}G(\ddot{x} \mid x)]^{-1}$ yields

$$\partial \rho_0 / \partial d_j = D_\rho \int \nabla_r \tau_\rho (\partial r / \partial d_j) \ \mathrm{d}G(\ddot{x} \mid x) = D_\rho \int \nabla_r \tau_\rho \ (\partial r / \partial d_j) \ g \ \mathrm{d}\nu,$$

where the second equality holds given B.3(b). (iv) A similar argument invoking B.3(a) and B.6(a) instead of B.3(b) and B.6(b) gives

$$\partial \mu_0 / \partial d_j = D_\mu \int [\tau_\mu (\partial \log g_d / \partial d_j) + \nabla_r \tau_\mu (\partial r / \partial d_j)] \ g_d \ \mathrm{d}\nu,$$

given the assumed existence of $D_\mu \equiv -[\int \nabla'_m \tau_\mu \ g_d \ \mathrm{d}\nu]^{-1}$. It follows that

$$\partial \mu_0 / \partial d_j - \partial \rho_0 / \partial d_j = D_\mu \int \tau_\mu (\partial \log g_d / \partial d_j) g_d \ \mathrm{d}\nu + \int [D_\mu \nabla_r \tau_\mu - D_\rho \nabla_r \tau_\rho g / g_d](\partial r / \partial d_j) g_d \ \mathrm{d}\nu.$$

The final expression of (iv) for $\delta_{0,j}$ holds by adding and subtracting terms appropriately. (v) The result follows immediately from the Minkowski and Cauchy-Schwarz inequalities. ■

8 References

Abadie, A., J. Angrist, and G. Imbens (2002), "Instrumental Variables Estimates of the Effects of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91-117.

Abadie, A. and G. Imbens (2002), "Simple and Bias-Corrected Matcing Estimators for Average Treatment Effects," NBER Technical Working Paper No. 283.

Ai, C. and Chen, X. (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795-1843.

Angrist, J. (1998) "Using Social Security Data on Military Applicants to Estimate the Effect of Voluntary Military Service on Earnings," *Econometrica*, 66, 249-288.

Angrist, J. and A. Krueger (1999), "Empirical Strategies in Labor Economics," in: O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Vol 3A. Amsterdam: Elsevier, pp. 1277 - 1368.

Bang-Jensen, J. and G. Gutin (2001). *Digraphs: Theory, Algorithms, and Applications.* London: Springer Verlag.

Barnow, B., G. Cain, and A. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," in E. Stromsdorfer and G. Farkas (eds.), *Evaluation Studies*, vol 5. San Francisco: Sage, pp. 43-59.

Bartle, R. (1966), *Elements of Integration.* New York: Wiley.

Berge, C. (1963), *Espaces Topologiques.* Paris: Dunod (translation by E.M. Patterson, *Topological Spaces.* Edinburgh: Oliver and Boyd).

Bettinger, E. and R. Slonin (2004), "Using Experimental Economics to Measure the Effect of a Natural Educational Experiment on Altruism," NBER Working Paper.

Billingsley, P. (1979). *Probability Theory and Measure.* New York: Wiley.

Blundell, R. and J. Powell (2003), "Endogeneity in Nonparametric and Semiparamtric Reression Models," in M. Dewatripoint, L. Hansen, and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eight World Congress*, vol II. New York: Cambridge University Press, pp. 312-357.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement.* Oxford: Oxford University Press.

Chalak, K. and H. White (2006), "An Extended Class of Instrumental Variables for the Estimation of Causal Effects," UCSD Department of Economics Discussion Paper.

Chen, X. (2005), "Large Sample Sieve Estimation of Semi-Nonparametric Models," New York University C.V. Starr Center Working Paper.

Chiang, A. (1984). *Fundamental Methods of Mathematical Economics.* New York: McGraw-Hill.

Dawid, A.P. (1979), "Conditional Independence in Statistical Theory," Journal of the Royal Statistical Society, Series B, 41, 1-31.

Dawid. A.P. (2000), "Causal Inference without Counterfactuals," *Journal of the American Statistical Association*, 95, 407-448.

Dawid, A.P. (2002), "Influence Diagrams for Causal Modeling and Inference," *International Statistical Review*, 70, 161-189.

Eckel, C., C. Johnson, and C. Montmarquette (2003), "Human Capital Investment by the Poor: Calibrating Policy with Laboratory Experiments," VPI Department of Economics Working Paper.

Engle, R., D. Hendry, and J.-F. Richard (1983), "Exogeneity," *Econometrica*, 51, 277-304.

Fershtman, C. and U. Gneezy (2001), "Discrimination in a Segmented Society: An Experimental Approach," *The Quarterly Journal of Economics*, 116, 351-377.

Firpo, S., N. Fortin, and T. Lemieux (2005), "Decomposing Wage Distributions: Estimation and Inference," UBC Department of Economics Working Paper.

Fisher, F. (1961), "On the Cost of Approximate Specification in Simultaneous Equation Estimation," *Econometrica*, 29, 139-170.

Fisher, F. (1966). *The Identification Problem in Econometrics.* New York: McGraw-Hill.

Fisher, F. (1970), "A Correspondence Principle for Simultaneous Equations Models," *Econometrica*, 38, 73-92.

Fisher, R.A. (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Gourieroux, C., A. Monfort, and A. Trognon (1984), "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681-700.

Grenander, U. (1981). *Abstract Inference.* New York: Wiley.

Hahn J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effect," *Econometrica*, 66, 315-331.

Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

Harrison, G., M. Lau, and M. Williams (2002), "Estimating Individual Discount Rates for Denmark: A Field Experiment," *American Economic Review*, 92, 1606-1617.

Heckman, J. (2005), "Econometric Causality," University of Chicago Department of Economics, manuscript.

Heckman, J. and J. Hotz (1989), "Alternative Methods for Evaluating the Impact of Training Programs," (with discussion), *Journal of the American Statistical Association*, 84, 862-874.

Heckman, J., H. Ichimura, and P. Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, 64, 605-654.

Heckman, J., H. Ichimura, and P. Todd (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261-294.

Heckman, J. and R. Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labor Market Data.* Cambridge: Cambridge University Press, pp. 156-245.

Heckman, J., J. Smith, and N. Clements (1997), "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487-535.

Heckman, J., S. Urzua, and E. Vytlacil (2005), "Understanding Instrumental Variables in Models with Essential Heterogeneity," University of Chicago Department of Economics, manuscript.

Heckman J. and E. Vytlacil (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669-738.

Hirano, K. and G. Imbens (2001), "Estimation of Causal Effects using Propensity Score Weighting: An Application to Right Heart Catheterization," *Health Services and Outcomes Research*, 2, 259-278.

Hirano, K. and G. Imbens (2004), "The Propensity Score with Continuous Treatments," in A. Gelman and X.-L. Meng (eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives.* New York: Wiley, pp. 73-84.

Hirano, K., G. Imbens, and G. Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161-1189.

Holland, P.W. (1986), "Statistics and Causal Inference" (with Discussion), *Journal of the American Statistical Asssociation*, 81, 945-970.

Hoover, K. (2001). *Causality in Macroeconomic*s. Cambridge: Cambridge University Press.

Hoover, K., (2004), "Lost Causes," *Journal of the History of Economic Thought*, 26, 149-164.

Imbens, G. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4-29.

Imbens, G. and W. Newey (2003), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," MIT Department of Economics Working Paper.

Karlan, D. (2005), "Using Experimental Economics to Measure Social Capital and Predict Finance Decisions," *American Economic Review*, 95, 1688-1699

Lechner, M. (1999), "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification," *Journal of Business and Economic Statistics*, 17, 74-90.

Lechner, M. (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independece Assumption," in M. Lechner and F. Pfeiffer (eds.), *Econometric Evaluations of Active Labor Market Policies in Europe*. Heidelberg: Physica-Springer, pp. 43-58.

Lehmann, E. (1974). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

List, J. (2004), "The Nature and Extent of Discrimination in the Market Place: Evidence from the Field," *Quarterly Journal of Economics*, 119, 49-89.

List, J. and D. Lucking-Reiley (2000), "Demand Reduction in Multi-Unit Auctions: Evidence from a Sportscard Field Experiment," *American Economic Review*, 90, 961-972.

List, J. and D. Lucking-Reiley (2002), "The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign," *Journal of Political Economy*, 90, 215-233.

Lucking-Reiley, D. (1999), "Using Field Experiments to Test Equivalence Between Auction Formats: Magic on the Internet," *American Economic Review*, 89, 1063-1080.

Matzkin, R. (2003), "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339-1375.

Matzkin, R. (2004), "Unobservable Instruments," Northwestern University Department of Economics Working Paper.

Matzkin, R. (2005), "Identification of Nonparametric Simultaneous Equations," Northwestern University Department of Economics Working Paper.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufman.

Pearl, J. (1993a), "Aspects of Graphical Methods Connected with Causality," in *Proceedings of the 49th Session of the International Statistical Institute*, pp. 391-401.

Pearl, J. (1993b), "Comment: Graphical Models, Causality, and Intervention," *Statistical Science*, 8, 266-269.

Pearl, J. (1995), "Causal Diagrams for Experimental Research" (with Discussion), *Biometrika*, 82, 669-710.

Pearl, J. (1998), "Graphs, Causality, and Structural Equation Models," *Sociological Methods and Research*, 27, 226-284.

Pearl, J. (2000). *Causality.* New York: Cambridge University Press.

Powell, J., J. Stock, and T. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403-1430.

Reichenbach, H. (1956). *The Direction of Time.* Berkeley: University of California Press.

Robins, J. (1989), "The Control of Confounding by Intermediate Variables," *Statistics in Medicine*, 8, 679-701.

Robins, J., S. Greenland, and F.-C. Hu (1999), "Estimation of the Causal Effect of a Time-Varying

Exposure on the Marginal Mean of a Repeated Binary Outcome," *Journal of the American Statistical Association*, 94, 687-712.

Robins, J., M. Hernan, and B. Brumback (2000), "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology*, 11, 550-560.

Robins, J., M. Hernan, and U. Siebert (2004), "Effects of Multiple Interventions," in M. Ezzati, A. Lopez, A. Rodgers, and C. Murray (eds.), *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, Vol 1. Geneva: World Health Organization, pp. 2191-2230.

Rosenbaum, P. (1984), "The Consequences of Adjustment for a Concomitant Variable that has been Affected by Treatment," *Journal of the Royal Statistical Society, Series A*, 147, 656-666.

Rosenbaum, P. (1987), "The Role of a Second Control Group in an Observational Study," (with discussion), *Statistical Science*, 2:3, 292-316.

Rosenbaum, P. and D. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. (1980), "Comment on 'Randomization Analysis of Experimental Data: The Fisher Randomization Test,' by D. Basu," *Journal of the American Statistical Association*, 75, 591-593.

Rubin D. (1986), "Which Ifs have Causal Answers?" (Comment on "Statistics and Causal Inference," by P. Holland), *Journal of the American Statistical Association*, 81, 961-962

Shipley, W. (2000). *Cause and Correlation in Biology: A User's Guide to Path Analaysis, Structural Equations, and Causal Inference.* Cambridge: Cambridge University Press.

Simon, H. (1953), "Causal Ordering and Identifiability," in W.C. Hood and T.C. Koopmans (eds.), *Studies in Econometric Methods.* New York: Wiley, pp. 49-74.

Spirtes, P., C. Glymour and R. Scheines (1993). *Causation, Prediction, and Search.* Cambridge, MA: MIT Press.

Stoker, T. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461-1481.

Strotz, R. and H. Wold (1960), "Recursive vs. Nonrecursive Systems: An Attempt at Synthesis," *Econometrica*, 28, 417-427.

Verma, T. and J. Pearl (1991), "Equivalence and Synthesis of Causal Models," in P.P. Bonissone, M. Henrion, L.N. Kanal, and J.F. Lemmer (eds.), *Uncertainty in Artificial Intelligence, vol. 6.*

Amsterdam: Elsevier, pp. 255-268.

White, H. (1980), "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 21, 149-170.

White, H. (1994). *Estimation, Inference and Specification Analysis.* New York: Cambridge University Press.

White, H. (2005a), "Time Series Estimation of the Effects of Natural Experiments," *Journal of Econometrics* (in press).

White, H. (2005b), *Causal, Predictive, and Explanatory Modeling in Economics.* Oxford: Oxford Univerity Press (forthcoming).

White, H. and K. Chalak (2006), "Parametric and Nonparametric Estimation of Covariate-Conditioned Average Causal Effects," UCSD Department of Economics Discussion Paper.

Wold, H. (1954), "Causality and Econometrics," *Econometrica*, 162-177.

Wold, H. (1956),"Causal Inference from Observational Data: A Review of Ends and Means," *Journal of the Royal Statistical Society, Series A*, 119, 28-50.

Wooldridge, J. (2002). *Econometric Analysis of Cross-Section and Panel Data.* Cambridge MA: MIT Press.

Wooldridge, J. (2005), "Violating Ignorability of Treatment by Controlling for Too Many Factors," *Econometric Theory*, 21, 1026-1029.